REMARKS

Claims 45, 51-52, and 70-73 have been amended. Support for these amendments can be found in Claim 47 as originally filed, and in the Specification on page 4, lines 9-13; page 9, line 32; page 17, lines 1-2 and 12-17; page 21, line 32 through page 22; line 8; page 26 lines 12-17 and 31-32 ; page 28, lines 7-9; page 29, lines 20-23; Example 7 and Tables 2 and 3. Claims 46 and 47 have been cancelled. New Claim 87 have been added. Support for the new claim can be found in Claim 73 as originally filed. Claim 49 has been canceled. Therefore, no new matter has been introduced by these amendments. The following addresses the substance of the Office Action.

Claim Objections

The Examiner has objected to Claim 85 for depending from Claim 44, which is drawn to a non-elected invention. Applicant has withdrawn Claim 85.

Compliance with 35 USC §112

Written description

The Examiner has rejected Claims 45-87 under 35 USC §112, first paragraph as failing to comply with written description requirement. More specifically, the Examiner believes that Claim 45 encompasses any and all manner of "discs", "registered data", "non-cleavable capture molecules", and "immobilized nucleic acids". Applicant disagrees.

> To satisfy the written description requirement, a patent application must describe the invention in sufficient detail that one of skill in the relevant art could conclude that the inventor was in possession of the claimed invention at the time the application was filed. *See Vas-Cath Inc. v. Mahurkar*, 935 F.2d 1555, 1563-64, (Fed. Cir. 1991). In view of the recent decision by the Federal Circuit, *Union Oil of California*, it is clear that an Applicant need not precisely recite each and every element of a claim limitation in the specification in order to satisfy the written description requirement. *See Union Oil of Cal. v. Atlantic Richfield Co.*, 208 F.3d 989 (Fed. Cir. 2000).

Applicant has now amended Claim 45 to specify that the disc is a compact-disc (CD), and the registered data is binary data concerning characteristics of the capture molecules or concerning treatment of a signal which results from binding between the target molecule(s) and the capture molecule(s). Support for this amendment can be found in Claims 47 and 49 as originally filed and in the Specification on page 9, line 32; page 26, lines 12-17, page 28, lines 7-9.

Claim 45 has been further amended to specify that capture and target molecules are selected from the group consisting of antibodies, proteins, receptors, ligands of said receptors, nucleic acids useful as diagnostic agents, nucleic acids useful for detecting presence of a pathogenic organism, and nucleic acid probes, wherein said nucleic acid probes are from nucleic acids encoding polypeptides, said polypeptides selected from the group consisting of enzymes, transcription factors, structural proteins, transporters, antibodies, antigens, receptors, markers of toxicity, bacterial markers, viral markers, oncogenes, tumor suppressors, senescence markers, tumor necrosis factors, proteins involved in apoptosis, inflammation, DNA damage and repair, oxidative stress, metabolism, and cell cycle. Support for these amendments can be found in the Specification as filed, for example, page 16, lines 19-27; page 17, lines 1-2 and 12-17; page 21, line 32 through page 22 line 8; page 26, lines 31-32; page 29, lines 20-23; Example 7 and Tables 2-3.

During the personal interview, the Examiner has raised concerns that Claim 45 may not satisfy 35 USC §101 utility requirements. The Examiner asserts that the capture and target molecules as recited in Claim 45 have no specified utility. The requirement of 35 U.S.C. 101 is that some specific, substantial, and credible use be set forth for the invention. Currently amended Claim 45 recites that capture and target molecules are selected from the group consisting of antibodies, proteins, receptors, ligands of said receptors, and nucleic acids encoding polypeptides, wherein said polypeptides are selected from the group consisting of enzymes, transcription factors, structural proteins, transporters, antibodies, antigens, receptors, markers of toxicity, oncogenes, tumor suppressors, senescence markers, tumor necrosis factors, proteins involved in apoptosis, inflammation, DNA damage and repair, oxidative stress, metabolism, and cell cycle, and therefore, have utility. Tables 1-3 in Specification provide some specific examples of the molecules that can be used as capture molecules on the disc of the invention. The sequences and functions of these molecules were known at the time the invention was made. Furthermore, EP 1 136 566 incorporated by reference in the Specification (page 7, lines 13-17) provides additional sequences of capture molecules. However, the Applicant asserts that because polynucleotides as a class share common chemical properties allowing any of them to be bound to the surface of the disc, and same is true to polypeptides as a class, ANY nucleic acid molecule, ANY receptor, ANY protein, ANY antibody raised against any protein, ANY polypeptide ligand

with known function can be attached to the surface of the disc of the invention. The 37 CFR 132 Declaration signed by a third party to be submitted shortly supports this statement.

The Applicant also provides a Declaration under 37 CFR 132 signed by Dr. Jose Remacle, the inventor of the present application. Applicants note that the claims are currently subject to an election of species in which the initial embodiment of the capture molecule to be searched is the embodiment where the capture molecule is a nucleic acid. However, should this embodiment be found to be allowable, Applicants note that in other embodiments the capture molecule may be selected from the group consisting of antigens, antibodies, receptors, ligands of receptors, receptor and enzyme peptides or a combination thereof. Accordingly, Applicants address all of these embodiments in the accompanying Inventor's Declaration.

As attested in the Inventor's Declaration, by December 30, 1997 those of skill in the art would appreciate the methodology set forth in the present application for fixing nucleic acids, peptides or polypeptides to the surface of the disc may be employed regardless of the sequence of such capture molecules.

For example, with respect to embodiments in which the capture molecule is a nucleic acid, the surface of the disc may be aminated as described, for example, on page 19, lines 15-31 of the Provisional application, as well as on page 20, lines 16-18, page 44, line 31-page 45, line 9, and in Example 1 of the present specification, thereby the nucleic acids may be bound to the amine groups on the disc as described at the foregoing locations of the Provisional and the present applications. As of December 30, 1997, those of skill in the art would appreciate that because the amino groups on the surface of the disc can be covalently bound to any nucleic acid regardless of its sequence, the methodology described in the specification is universally applicable to all nucleic acids. Accordingly, as of December 30, 1997 those skilled in the art would appreciate that the application contained sufficient description of how to bind any desired nucleic acid to the surface of the disc.

Other methodology for fixing capture molecules bearing amino groups to the surface of the disc is described in the present specification at page 20, lines 7-16. As of December 30, 1997, those of skill in the art would appreciate that because this methodology will work with any capture molecule bearing an amino group, the application contained sufficient description of how to bind any desired capture molecule bearing an amino group to the surface of the disc (see, for

example, page 19, lines 15-31 of the Provisional application and Rasmussen et al. 1991 "Covalent immobilization of DNA onto polystyrene microwells: the molecules are only bound at the 5" end" *Anal. Biochem.* **198**:138-142, which is referenced in the Provisional application on page 24, line 32). Thus, as of December 30, 1997 those skilled in the art would appreciate that the application contained sufficient description of how to bind any desired molecule bearing an amino group to the surface of the disc.

Other methods for fixing any desired capture molecule bearing a reactive group by deprotecting or protecting the reactive group or by synthesizing the capture molecule on the surface of the disc are described in the present specification at page 30, line 19 - page 31, line 2.

Therefore, Claim 45 as amended is supported by the Specification as filed and complies with 35 USC §101 and §112, first paragraph.

The Examiner expressed concerns regarding the ability of one skilled in the art to convert an output data (digital information, as electronic string of 1' and 0') into any desired form be it words, numbers, notes, etc. In the submitted 132 Declaration signed by the inventor it is stated that as of December 30, 1997, those skilled in the art were familiar with how to convert digital information on the disc into a desired form of output, such as words, notes, numbers etc. As described in the Provisional application at page 10, lines 13-32, and in the present specification at page 4, lines 1-35, and in the conventional CD technology available on December 30, 1997, data is stored on the CD as pits and lands. The pits and lands are converted into digital data (1's and 0's) when read by a laser. Specifically, each pit is converted into a binary 1 and each land is converted into a binary 0. As of December 30, 1997, CD's were being utilized to provide output in a variety of formats and those skilled in the art would appreciate that such technology was standard. In fact, as demonstrated in the attached History of CD Technology (Exhibit 1of the Inventor's Declaration) CD technology was quite mature at the time the present application was filed. Thus, as of December 30, 1997, those skilled in the art could readily convert binary digital information into any desired form of output.

Similarly, as of December 30, 1997, those skilled in the art were familiar with the use of lasers to read data from a disc (see Provisional application at page 9, line 12 through page 10, line 23, page 11, line 31 - through page 12, line 29). As discussed in the preceding paragraph, data present on the disc is converted into 1's or 0's using conventional laser technology. The

presently claimed discs contain registered data stored on the disc as conventional pits or lands. When the laser light shines on a pit, a signal transition is generated which is converted into a binary 1 (see present specification at page 11, lines 17-31). Likewise, when the laser light shines on a land, no signal transition is generated and this is converted into a binary 0. Again, this technology is the standard technology which was conventionally used as of December 30, 1997 to retrieve data from a standard CD at the time the present application was filed.

The presently claimed discs also generate binary data reflecting the binding of a target to a capture molecule. In some embodiments, binding of a target to a capture molecule results in formation of a precipitate, which forms a mound on the surface of the disc. The mound perturbs the laser reflection (just as a pit perturbs the laser reflection) and is converted into a binary 1 (see Provisional application, page 14, line 18 through page 16, line 24, page 21, lines 20-23, and Figure 3; and the present specification, page 23, line 24-page 24, line 33) Thus, binding of a target to a capture molecule is detected using the standard laser technology used to read data from a conventional CD which was available as of December 30, 1997.

As stated in the Declaration signed by the Inventor, Jose Remacle, by December 1997, those skilled in the art were also quite familiar with technology which can be used to quantitate a signal resulting from the binding of a target molecule to a capture molecule on the surface of the disc.

As of December 30, 1997, those of skill in the art were familiar with technology which can be used for quantitating the signal resulting from binding of a target to a capture molecule on the surface of the claimed discs. For example, binding of a target to a capture molecule can be quantitated using standard laser technology employed in conventional CD players which were available as of December 30, 1997, standard fluorescence reading technologies available as of December 30, 1997, and image recognition software available as of December 30, 1997 (see Provisional application, page 3, line 6 - page 4, line 12, and page 24, lines 8-15).

The present specification also describes such technologies In addition, the present specification describes quantitation at pages 18-23, page 45, lines 10-2. In addition, Figures 17 and 20 provide actual quantitation curves obtained with the presently claimed discs.

The Examiner expressed concerns regarding whether the specific components of a kit of Claims 73 and 86 are well known in the art. The Inventor's Declaration submitted herewith

attests that as of December 30, 1997 one of skill in the art was familiar with reactants which can be used to bind a target molecule in a sample to a capture molecule on the disc of the present invention. In fact, the Declaration by Jose Remacle provides Exhibit 2 which shows that as of December 30, 1997 a variety of DNA chips were available, and that those skilled in the art were familiar with buffers and hybridization conditions used with such chips.

In addition, as of December 30, 1997, those skilled in the art were familiar with reactants which can be used to detect binding between a target molecule and a sample. In particular, a variety of reactants which could be used to form precipitates at a bound target were known as of December 30, 1997. For example, reactants for generating a precipitate by silver staining, fluorescent reagents, and colorimetric reactants were known to those of skill as of December 30, 1997 (see Provisional application at page 3, line 30 - page 4, line 6, and page 22, line 9 - page 24, line 7). In addition, the present specification provides numerous examples of such reactants at page 17, line 28-page 18, line 26, page 22, line 4-page 25, line 8 and Examples 1-10, and Examples 13-15.

Enablement

The Examiner has rejected Claims 45-87 under 35 USC §112, first paragraph, as failing to comply with enablement requirement. More specifically, the Examiner believes that the Specification is not enabling as to nucleic acids immobilized on the surface of a disc, the reader and the handler. The Examiner further cited that "...when there is no disclosure of any specific starting material or any of the conditions under which a process can be carried out, undue experimentation is required; there is a failure to meet the enablement requirement that cannot be rectified by asserting that all the disclosure related to the process is within the skill of the art" Applicant respectfully disagrees. The rule according to MPEP 2164 is:

> The purpose of the requirement that the specification describe the invention in such terms that one skilled in the art can make and use the claimed invention is to ensure that the invention is communicated to the interested public in a meaningful way. The information contained in the disclosure of an application must be sufficient to inform those skilled in the relevant art how to both make and use the claimed invention. However, to comply with 35 U.S.C. 112, first paragraph, it is not necessary to "enable one of ordinary skill in the art to make and use a perfected, commercially viable embodiment absent a claim limitation to that effect. The test of enablement is whether one reasonably skilled in the art could make or use the invention from the disclosures in the patent

coupled with information known in the art without undue experimentation.. A patent need not teach, and preferably omits, what is well known in the art.

Here, for the Examiner's convenience, Applicant provides the exact place in the Specification where the starting materials are disclosed:

Claim	Support in the Specification
45. A compact-disc (CD) comprising:	3:19-4:7;
one or more capture molecule(s) bound to a specific surface area of said disc,	
wherein said capture molecule(s) do(es) not comprise a cleavable spacer,	17:1-3
said capture molecule(s) providing a site for specific binding with one or more target molecule(s) to be detected, identified and/or quantified,	17:3-17
wherein said capture and target molecules are selected from the group consisting of antibodies, proteins, receptors, ligands of said receptors, nucleic acids useful as diagnostic agents, nucleic acids useful for detecting presence of a pathogenic organism, and nucleic acid probes, wherein said nucleic acid probes are from nucleic acids encoding polypeptides,	16:23-27, 17:12-17; 26:31-32; 29:20-23; 48:1-2; Example 7, and Claim 46
said polypeptides selected from the group consisting of enzymes, transcription factors, structural proteins, transporters, antibodies, antigens, receptors, markers of toxicity, oncogenes, tumor suppressors, senescence markers, tumor necrosis factors, proteins involved in apoptosis, inflammation, DNA damage and repair, oxidative stress, metabolism, and cell cycle; and	Tables 2 and 3
registered data concerning characteristics of the capture molecules or concerning treatment of a signal which results from binding between the target molecule(s) and the capture molecule(s),	4:9-13; 4:31-5:7
wherein said registered data is binary data.	4:14-30; 5:10-21

Furthermore, Tables 2 and 3 provide a list of specific nucleic acid sequences that can be immobilized on the surface of a disc. 58 out of 59 listed sequences were publicly available before December 30, 1997 (see Exhibit B). As the Applicant asserted before and as stated in the presently filed 132 Declarartion by Jose Remacle, and 132 Declarartion by the Third Party (to be submitted shortly) ALL nucleic acids share common chemical properties allowing them to be bound to the surface of a compact disc using the methodology set forth in the Specification. Moreover, the Specification provides examples of detecting of a specific DNA hybridized to

target molecules on the disc (Examples 1 and 7) as well as some exemplary sequences which could be bound to the disc (Tables 2 and 3) and provides the meaning of the term nucleic acid, oligonucleotide etc. on page 7, lines 13-17. In view of the fact that all nucleic acids possess the same chemical moieties which can be used to attach them to the surface of the disc, it is not necessary for Applicants to recite all the sequence of every nucleic acid which can be bound to the disc. As of December 1997, more than 1.5 million DNA/RNA and polypeptide sequences were publicly available (Exhibit C), and since then this number grew to more than 40 million (Exhibit D). As stated in the 35 USC 1.132 Declaration by the inventor, as well as in the Declaration by the Third party, just as a skilled artisan can appreciate that any desired nucleic acid could be used as a probe on a nitrocellulose filter, those skilled in the art would similarly appreciate that any desired nucleic acid could be bound to the surface of the disc. Furthermore, the methodology set forth in the specification allows the skilled artisan to fix any desired nucleic acid to the surface of the disc.

Furthermore, the rule according to MPEP 2164 is:

> To overcome a prima facie case of lack of enablement, applicant must demonstrate by argument and/or evidence that the disclosure, as filed, would have enabled the claimed invention for one skilled in the art at the time of filing. This does not preclude applicant from providing a declaration after the filing date which demonstrates that the claimed invention works. Such a showing also must be commensurate with the scope of the claimed invention, i.e., must bear a reasonable correlation to the scope of the claimed invention.

During the personal interview, the Applicant's representative provided the Eppendorf Array Technologies Bio-CD color printout from http://www.eppendorf.com/eat/tech/bio_cd.html which disclosed the now commercially available claimed invention, therefore demonstrating that the claimed invention works.

Therefore, Applicant respectfully asserts that the Claims 45-87 as currently amended are enabled and their rejection should be withdrawn.

Request for additional information

The Examiner had previously requested additional information as to the differences between the present CIP application and US 60/071,726. Applicants' representatives note that it was not possible to generate an informative electronic comparison between the provisional application and the present CIP application using the software available to us. Hence, we were

unable to do a word by word comparison of the two documents. However, we note the following differences between the present CIP application and US 60/071,726 are as follows.

-Fixation of capture molecules upon the CD surface according to an array;

-fixation of capture molecules upon the CD surface area(s) opposite to the surface area(s) comprising the registered data;

-expanded discussion on embodiments having chambers, cavities and microchannels upon the CD surface;

-description of a CD platform;

-expanded discussion on the term "registered data"; and

-Examples 1-16 and Figures 3-20 are new compared to the provisional application.

CONCLUSION

Applicants have endeavored to address all of the Examiner's concerns as expressed in the outstanding Office Action. Accordingly, amendments to the claims, the reasons therefor, and arguments in support of the patentability of the pending claim set are presented above. In light of the above amendments and remarks, reconsideration and withdrawal of the outstanding rejections is specifically requested. If the Examiner finds any remaining impediment to the prompt allowance of these claims that could be clarified with a telephone conference, the Examiner is respectfully requested to initiate the same with the undersigned.

Please charge any additional fees, including any fees for additional extension of time, or credit overpayment to Deposit Account No. 11-1410.

Respectfully submitted,

KNOBBE, MARTENS, OLSON & BEAR, LLP

Dated: May 23, 2005

By: _____

Marina L. Gordey
Registration No. 52,950
Agent of Record
Customer No. 20,995
(805) 547-5580

1664640
041805

CURRICULUM VITAE

REMACLE, José, A. L.

ADDRESS *Home* Chemin des Pierres 14 5730 Malonne, Belgique
Tél. : 32-(0)81-44.10.08

Business Facultés Universitaires Notre-Dame de la Paix
Rue de Bruxelles 61 5000 NAMUR Belgique
Tel : 32-(0)81-724123 (Office) Fax : 32-81-724135
Email : jrem@biocell.fundp.ac.be

PERSONAL INFORMATION :

Place and date of birth : : Nassogne (Belgium), on the 31st August, 1946
Marital status : married
Nationality Belgian

DEGREES

- *Bachelor of Chemistry with maxima cum laude, 1970*
  Université Catholique de Louvain, Belgium.
- *Ph.D. in Sciences, Biochemistry, with maxima cum laude, 1973*
  Université Catholique de Louvain, Belgium.
  Directeurs de thèse : Profs H. Beaufay and A. Trouet.
  Laboratoire de Chimie Physiologique, Prof. C. de Duve.

POSITIONS

1970 - 1971 : *Junior research of the National Fondation for Scientific Research (F.N.R.S.)*
1971 - 1974 : *Research assistant of the F.N.R.S.*
1973 - 1974 : Fellowship of "Belgian American Educational Foundation" (Bourse C.R.B.)
1974 : Research fellowship of the European Molecular Biology Organization (E.M.B.O.)
1974 *Associate professor* Facultés Universitaires Notre-Dame de la Paix, Namur.
1980: *Professor* Facultés Universitaires Notre-Dame de la Paix, Namur
Director of the Laboratory of Cellular Biochemistry.
1985: *Full Professor, with tenure*
1992: *Visiting Scientist* University of Maryland, Baltimore County Campus

AWARDS

1968 : Prix de "l'Union Carbide European Research Associates"
1973 : Bourse William Hallam Tuck, of the fondation Francqui
1984 : Prix Vander Stricht de la Fondation André Vander Stricht.
1992 : Senior Research Scholar at the University of Maryland, Baltimore for 1992-1995.

PROFESSIONAL EXPERIENCE

Research stage at the Rockefeler University, Prof. C. de Duve, from July to September 1973.

Post-doctoral research at the University of California, San Diego, U.S.A., from September 1973 to August 1974, in the laboratory of Prof. S. J. Singer.

Scientific mission of 4 months at the Biochemical Engineering Department of the University of Maryland in Baltimore,Laboratory of Prof G. Rao, in 1992.

Scientific mission at the Biochemical Engineering Department of the University of Maryland at Baltimore as Senior Research Scolar in March-April 1993.

SCIENTIFIC RESPONSABILITIES :

Head of the laboratory of cellular biochemistry

Actual composition (1997)
6 PhDs in Science full research
12 PhDs Students
6 Graduate full research
5 Under-graduate students
6 technicians

Students and researchers already formed
Director of 16 PhDs Thesis passed from 1981 to 1997.
Director of 74 graduate students from 1974 to 1997.

RESEARCH CONTRACTS

Reseach Contracts with Industries

30 research contracts with Laboratoires Dausse, Synthélabo, Solvay-Biotec, Compagnie des développements agro-alimentaires (CDA), Kali-Chemie Pharma, La Floridienne, CELAC, laboratoires Oberval, Laboratoires Beaufour, UCB-Pharma, Lambdatech, Lipha, IPSEN, Zyma, Madaus Pharma, Servier.

Scientific Grants and Research contracts

14 contracts with the FNRS, FRFC,IRSIA, and Région Wallonne

PROFESSIONAL AND SCIENTIFIC ASSOCIATIONS

Member of 15 scientific societies
Member of 33 Ph.D. thesis jurys
Member of the research comitte for Biomed 1 and 2 of the EEC

PRESENTATIONS OF RESULTS IN SCIENTIFIC CONGRES

171.presentations in scientific meetings as author or co-author.

INVITED OR PLENARY CONFERENCES OR LECTURES.

75 presentations in scientific meetings under invitation

MAIN SUBJECTS OF RESEARCH

Cellular Ageing and modelisation of the ageing process using the thermodynamics of open systems; the role of Free radical and the importance of the antioxidant enzymes.

Study of endothelial cells under hypoxia in correlation with the development of varicose diseases

Development of new diagnostic assays using bioluminescence: ELISA, DNA probes for virus and bacteria detection.

PUBLICATIONS

The author's scientific output consists of 125 research papers in peer-reviewed international journals

History of CD Technology

History of OneOff

Mission Statement

History of CD Technology

1841 Augustin-Louis Cauchy Proposes a Sampling Theorem.

1842 Charles Babbage Proposes analytical engine for performing and storing calculations.

1854 George Boole publishes "An Investigation Into the Laws of Thought." A book that contained, among other things, theories that were later used to build digital circuits.

1855 Leon Scott de Martinville invents the phonoautograph, a machine that records vibrations on a carbonized paper cylinder.

1876 Alexander Graham Bell introduces the telephone

1877 Thomas Edison invents the phonograph while trying to invent a device that would record and repeat telegraphic signals (digital)

1887 Emily Berliner replaces Edison's wax cylinder phonograph with the audio disc.

1915 78 R.P.M records introduced

1922 J.R. Carson examines the idea of time sampling for communications

1928 Harry Nyquiest publishes "Certain Topics in Telegraph Transmission Theory." His theory contained proof that the technology used in todays audio cd's could work.
33 1/3 Records Introduced

1937 A. Reeves invents pulse code modulation (PCM), a technology used by computers and CD's for audio in the present day.
H. Aiken from Harvard approaches IBM and proposes a electrical computing machine.

1943 The U.S. Army turns on the first computer (ENIAC) at the University of Pennsylvania.

1947 Magnetic Tape Recorders hit the U.S. market.

1948 The transistor is invented by Bell Laboratories.
Claude E. Shannon publishes "A Mathematical Theory of Communication." -- Yet another important development for theories used in CD technology

1949 45 rpm records hit the U.S. market, thanks to microgroove technology.

1950 Richard W. Hamming publishes information about error detection/correction codes. It would be impossible for CD's to work without error correction.

1958 Invention of the Laser.
Stereo LP's produced.
Integrated Circuit introduced by Texas Instruments

1960
Computer Music experiments take place at major laboratories.
I.S. Reed and G. Soloman publish information on multiple error correction codes. These come to be known as the "Reed-Solomon"

Codes which are the codes used for enconding and reading CD's.
Working Laser produced.

1967 NHK Technical Research Institute demonstrates a 12-bit PCM digital audio recorder with a 30 kHz (30,000 times per second) sampling rate. The digital recording goes onto a high-grade video tape.

1969 Sony introduces it's 13-bit PCM digital recorder at a 47.25 kHz (47,250 time per second) sampling rate. The digital recording is sent to a 2" video tape.
Klass Compaan, a Dutch physicist comes up with the idea for the Compact Disc.

1970 At Philips, Compaan and Pete Kramer complete a glass disc prototype and determine that a laser will be needed to read the information.

1971 Microprocessor produced by Intel
Digital Delay line used by BBC's studios (first digital audio device).

1972 Compaan and Kramer produce color prototype of this new compact disc technology

1973 BBC and other broadcast companies start installing digital recorders for master recordings.

1977 Mitsubishi, Hitachi & Sony show digital audio disc prototypes at the Tokyo Audio Fair.
JVC Develops Digital Audio Process

1978 Philips releases the video disc player
Sony sells the PCM-1600 and PCM-1 (digital audio processors)
"Digital Audio Disc Convention" Held in Tokyo, Japan with 35 different manufacturers.
Philips proposes that a worldwide standard be set.
Polygram (division of Philips) determined that polycarbonate would be the best material for the CD.
Decision made for data on a CD to start on the inside and spiral towards the outer edge.
Disc diameter originally set at 115mm.
Type of laser selected for CD Players.

1979 Prototype CD System demonstrated in Europe and Japan.
Sony agrees to join in collaboration.
Sony & Philips compromise on the standard sampling rate of a CD -- 44.1 kHz (44,100 samples per second)
Philips accepts Sony's proposal for 16-bit audio.
Reed-Solomon code adopted after Sony's suggestion.
Maximum playing time decided to be slighty more that 74 minutes.
Disc diameter changed to 120mm to allow for 74 minutes of 16-bit stereo sound with a sample rate of 44.1 kHz

1980 Compact Disc standard proposed by Philips & Sony.

1981 Matsushita accepts Compact Disc Standard
Digital Audio Disc Committee also accepts Compact Disc Standard.
Sharp achieves production of semiconductor laser.
Philips & Sony collaboration ends.

1982 Sony & Philips both have product ready to go.
Compact Disc Technology is introduced to Europe and Japan in the fall.

1983 Compact Disc Technology is introduced in the United States in the

spring
The Compact Disc Group formed to help market.
CD-ROM Protoypes shown to public
30,000 Players sold in the U.S.
800,000 CD's sold in the U.S.

1984 Second Generation & Car CD players introducted.
First Mass Replication Plant in the United States built.
Portable (i.e., Sony DiscMan) CD Players sold.

1985 Third generation CD Players released.
CD-ROM drives hit the computer market.

1986 CD-I (Interactive CD) concept created.
3 Million Players sold in U.S.
53 Million CD's sold in U.S.

1987 Video CD format created.
Allen Adkins of Optical Media International joins with SonoPress in Amsterdam and demonstrates a desktop system for pre-mastering CD's (Adkins and SonoPress, produced a replicated CD in less than 24-hours using this system).

1988 CD-Recordable Disc/Recorder Technology Introduced

1990 28% of all U.S. households have CD's.
9.2 million players sold annually in the United States.
288 million CD's sold annually in the United States.
World Sales close to 1 Billion

1991 CD-I format acheived.
CD-Recordable Introduced to the Market
"QuickTopix" the first CD-R pre-mastering Software introduced by Allen Adkins.

1992 CD-R Sales reach 200,000

1996 DVD Technology Introduced.
Prices of Recorders and CD-R Media go down significantly.
High Demands cause World-Wide CD-R Media Shortage.

1997 DVD Released.
DVD Players/Movies hit consumer market.
DVD-R standard created (3.9 Gig).
Mitsui builds it's first CD-R production plant in the U.S.
World-wide shortage ends.
Price of CD-R media lower than ever imagined.

1998 DVD-RAM, DVD-Recordable systems/equipment hits market.
DVD-Video/ROM authoring tools hits the market.
CD-R prices continue to drop.

1999 DVD-Video Becomes main stream.
Consumers begin purchasing DVD Players & Movies on a mass level.
Most major film studios have titles on DVD.
DIVX Dies (DIgital Video eXpress).
Second Generation DVD Burners.
4.7 Gig DVD-R Media Developed.

Source

1841-1991 **1991-1999**

Pohlmann, Ken C.
"The Compact Disc Handbook, 2nd Edition" (Click to See this Book at Amazon.Com)
Copyright 1992 & 1989 A-R Editions, Inc.

DNA chips: State-of-the art

Graham Ramsay

Wolpert Polymers, Inc., 4308 Reedy Avenue, Richmond, VA 23225-4636 (e-mail ramsayg@aol.com).

Received 10 November 1997; accepted 4 December 1997

The technology and applications of microarrays of immobilized DNA or oligonucleotides are reviewed. DNA arrays are fabricated by high-speed robotics on glass or nylon substrates, for which labeled probes are used to determine complementary binding allowing massively parallel gene expression and gene discovery studies. Oligonucleotide microarrays are fabricated either by in situ light-directed combinatorial synthesis or by conventional synthesis followed by immobilization on glass substrates. Sample DNA is amplified by the polymerase chain reaction (PCR), and a fluorescent label is inserted and hybridized to the microarray. This technology has been successfully applied to the simultaneous expression of many thousands of genes and to large-scale gene discovery, as well as polymorphism screening and mapping of genomic DNA clones.

Keywords: oligonucleotide array, cDNA array, gene discovery, gene expression, sequencing by hybridization

This review describes recently developed DNA chip technology and applications to gene discovery and expression, detection of mutations or polymorphisms and mapping[1–3]. No de novo sequencing studies were reported.

Two variants of the chip exist: In Format I DNA is immobilized to a solid surface such as glass and exposed to a set of labeled probes either separately[4] or in a mixture[5]; in Format II, an array of oligonucleotide probes is synthesized either in situ or by conventional synthesis followed by on-chip immobilization[6]. The array is exposed to labeled sample DNA, hybridized, and complementary sequences are determined (Fig. 1).

DNA chip structure and operating principles

DNA chip design. In complementary DNA (cDNA) chips immobilized targets of single-stranded cDNA are hybridized to ssDNA fluorescent probes produced from total mRNAs, to monitor expression levels of target genes. For example, cDNA from *Arabidopsis thaliana*[7,8] was robotically printed on glass microscope slides coated with poly-L-lysine and denatured. Fluorescently-labeled probes were labeled with fluorescein or lissamine and hybridized to the array under stringent conditions. Human acetylcholine receptor (AChR) mRNA was included as an internal standard. Hybridization was measured with a laser scanner and displayed as a pseudocolor display of differential expression.

Shalon et al. extended this technology by producing microarrays of genomic DNA from λ clones of *Saccharomyces cerevisiae*[9]. Hybridization and differential expression analysis was performed as above.

Very large-scale cDNA microarrays. Drmanac et al. have developed methods for the production of cDNA and genomic DNA microarrays containing up to 31,104 cDNA clones on comparatively large, single nylon membranes for gene expression and discovery experiments[4–6]. PCR protocols were designed to amplify several thousand clones per day. Target DNA was spotted onto replicate membranes in an automated procedure[10]. Membranes were hybridized sequentially with 200–320 short oligonucleotide probes labeled with ^{32}P, washed, and imaged on a storage phosphor screen. Hybridization data were processed to group identical genes into clusters and assign expression levels to each.

DNA chip production by in situ synthesis of oligonucleotides. The GeneChip is produced by adapting semiconductor photolithography to synthesize oligonucleotide probe sequences in situ on a glass substrate one cm square[11–13]. Conventional phosphoramidite-based DNA synthesis technology was modified by use of a photolabile protecting group on the terminal hydroxy group of a spacer group linked to the substrate surface (Fig. 2). The first step of the synthesis was removal of the protecting group and generation of free hydroxyl groups at positions illuminated by UV light shining through a photolithographic mask. In the second step, 5′-protected phosphoramidite was added to deprotected sites. Following capping of unreacted hydroxyls, oxidation and washing a second mask was used to synthesize the next nucleotide at the requisite probe sites (Fig. 3). The process was repeated until the required set of oligonucleotides had been synthesized. The use of combinatorial masking strategies generates high-density microarrays using comparatively few synthesis steps and the number of probes increases exponentially with a linear increase in the number of synthesis cycles. For a given oligonucleotide containing n nucleotides there are 4^n possible structures which can be produced in $4 \times n$ chemical steps. For example a complete set of octanucleotides consists of 65,536 probes and can be produced in 32 chemical steps in 8 hours.

Target DNA hybridization and epifluorescence detection. Hacia et al. fabricated GeneChips containing 96,600 probes per chip[14]. The DNA sample was labeled with phycoerythrin, hybridized to the chip and was scanned with a confocal epifluorescence microscope. To minimize experimental errors a reference target labeled with fluorescein was included in the sample. The ratio of reference to target fluorescence was used as a measure of hybridization efficiency. This procedure normalized the different responses obtained between stronger G+C base pairing and weaker A+T binding. The fluorescent image was digitized and hybridization displays were produced by computer.

Photolithography has been used to define individual elements of 5–10 μm, which corresponds to an array density of the order of 10^6 probes cm^{-2} (ref. 15). Diffraction of light at the mask apertures limits the element size using photo-labile groups but semiconductor technology using photoresists routinely produces submicron structures and this technology has been adapted to chip synthesis[11]. McGall and co-workers found it necessary to protect the growing oligonucleotide chains from acid-catalyzed deprotection with a layer of polyimide under the photoresist (Fig. 4). Initial work gave features of 8 μm size, which corresponds to a theoretical probe den-

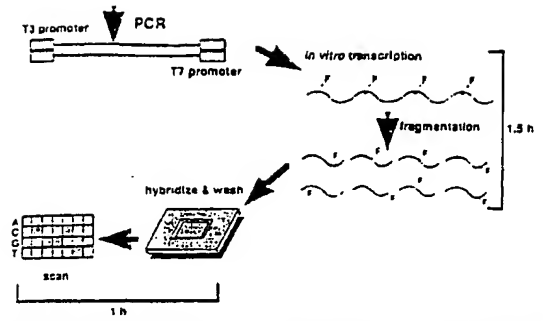


Figure 1. Sample preparation and hybridization to oligonucleotide array. (Reprinted with permission from *BioTechniques*, 1995, 19(3):445. Copyright 1995, Eaton Publishing.).

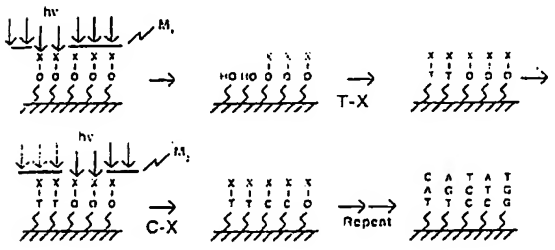
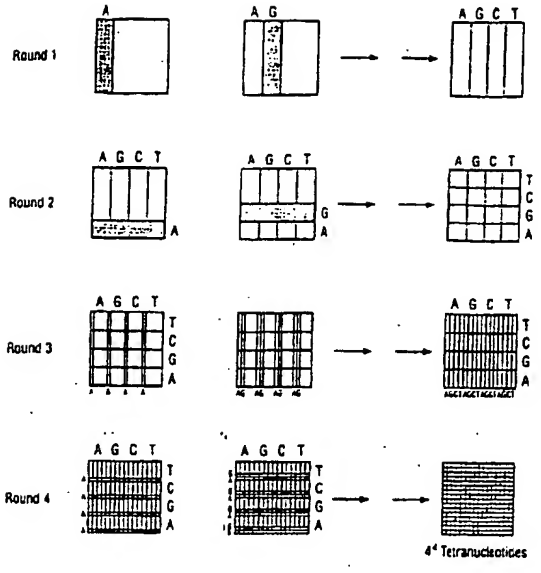


Figure 2. Light-directed synthesis of oligonucleotides. (Reprinted with permission from *Proc. Nat. Acad. Sci. USA*, 1994, 91:5023. Copyright 1994, National Academy of Sciences, USA).



Probe length	Chemical steps	Number of possible probes
4	16	256
8	32	65,536
10	40	1,048,576
15	60	1,073,741,824

Figure 3. Combinatorial array synthesis. (Reprinted with permission from *Chemtech*, February 1997, 27(2):27. Copyright 1997, American Chemical Society).

sity of the order of 10^6 probes cm^{-2} and it is expected that optimization will permit features of approximately 1 μm.

Oligonucleotide arrays produced by off-chip DNA synthesis. Yershov et al.[36] have produced DNA chips by conventional synthesis of DNA followed by robotic immobilization gel-coated glass plates at a density of 20,000 to 30,000 different probes cm^{-2}. The gel permits high oligonucleotide loading and enhanced hybridization but only short probes can diffuse into it. Addition of soluble probes extends the utility of these microarrays.

Direct electric field control to determine single base mutations in DNA. Heller et al. used electric fields to greatly accelerate the hybridization of labeled probe to immobilized sample oligonucleotides[37]. The microarray was fabricated on a silicon chip one centimeter square (Fig. 3). The silicon substrate was thermally oxidized and then platinized to form a 1 mm × 1 mm array of 25 microelectrodes. A layer of silicon nitride was deposited and etched into a sample well over each electrode. The electrodes were covered by a permeation layer of streptavidin-derivatized agarose to which biotinylated DNA sample was coupled under a positive potential.

Applications

Use of DNA arrays for gene expression and discovery. DNA chips have been used to measure expression levels of genes in plant[38], yeast[39,40] and human[41–43] samples. In initial work Schena et al. measured differential expression in *Arabidopsis thaliana* by using a microarray of 48 duplicate cDNA elements simultaneously assayed with a mixed set of probes labeled with fluorescein or lissamine. Negative controls, rat glucocorticoid receptor, and yeast TRP4 cDNAs were included. The expression of the bound probes was calibrated against the response of spiked human AChR mRNA. There were no detectable response from the rat and yeast controls, showing that the specificity of the array was high. The limit of detection was measured at 1:50,000 w/w of total mRNA. Target structures were assigned by sequencing and comparison to the National Center for Biotechnology Information database, resulting in the matching of 45 of the 48 cDNAs.

The transgenic plant expressed the HAT4 gene at a level 50 times that of the wild-type and the other 44 genes gave expression levels within a factor of five between the two samples. Control probes of fluorescein-labeled rat glucocorticoid receptor cDNA and lissamine-tagged yeast TRP4 cDNA showed that there was no cross-talk between the two fluors. This experiment demonstrated that expression levels of two biological samples could be determined by use of the same array and by simultaneous exposure to two sets of probes.

Subsequently, probes from *Arabidopsis* root and leaf mRNA were used to explore differences in expression. The light-regulated CAB1 gene was expressed in leaf tissue at a level 500 times that in root, as expected, because it is known to be highly repressed in root cells. Twenty-six other genes differed by more than a factor of five.

Schena et al. monitored the expression of 1046 human cDNAs of unknown sequence using two-color differential expression analysis of heat shock or phorbol ester-regulated genes[44]. This system was ten times more sensitive than used previously and gave a detection limit of 1:500,000 w/w total human mRNA. The effects of heat shock or exposure to phorbol ester on human T cells were examined on the same microarray by exposure to probes derived from total mRNA from cells grown under these conditions. The cDNAs of induced genes were sequenced and identified by comparison to known structures. Heat shock resulted in the induction of known heat shock genes for molecular chaperones and mediators of molecular degradation. Similarly, phorbol ester exposure result-

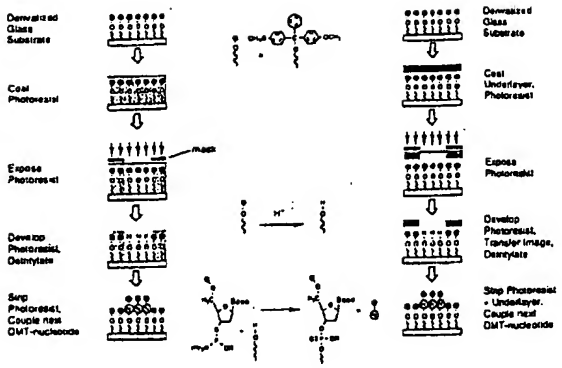


Figure 4. Oligonucleotide fabrication using polymeric photoresists. (A) Single layer process (B) Bilayer process using an inert polymer underlayer to protect surface oligonucleotide chemistry. (Reprinted with permission from *Proc. Nat. Acad. Sci. USA*, 1996, 93:13557. Copyright 1996, National Academy of Sciences, USA).

Figure 5. Cross-section of a Nanogen DNA chip. (Reprinted with permission from *Proc. Nat. Acad. Sci. USA*, 1997, 94:1120. Copyright 1997 National Academy of Sciences, USA).

ed in the detection of genes characteristic of the phorbol ester signaling pathway such as phosphatase of activated cells and nuclear factor-κB1. Additionally, three known genes that were expressed at low levels had not been previously assigned to this pathway. Four novel genes were identified, each expressed at a low level. It is likely that previous conventional screens were not sufficiently sensitive for their detection. These experiments demonstrated the ability of cDNA arrays to rapidly provide data for correlation of gene expression to biochemical pathways.

This technology was extended to the study of gene expression characteristic of the inflammatory diseases rheumatoid arthritis (RA) and inflammatory bowel disease (IBD)[20]. Probes were produced from RA tissue or IBD mucosa labeled with either Cyt 3 or Cyt 5 fluors and exposed to a microarray consisting of cDNA targets from genes known to be involved in the disease processes. Genes known to be expressed in inflammatory diseases were observed, such as tumor necrosis factor, interleukins and granulocyte colony-stimulating factor. Some genes were expressed that had not previously been associated with inflammatory diseases, such as human matrix metallo-elastase (HME) and melanoma growth stimulatory factor. The newly assigned genes could become therapeutic targets. Differential expression between the two disease states was effected by use of a microarray of 1046 cDNA clones from a peripheral blood library. Tissue inhibitor of metalloproteinase 1, ferritin light chain and manganese superoxide dismutase were more strongly expressed in RA tissue, demonstrating the ability of cDNA chips to rapidly provide information about the genetic basis of disease.

DeRisi et al. investigated the genetic basis of tumorgenicity by the use of two color fluorescence[21]. The tumorigenic properties of human melanoma cell line UACC-903 can be reversed by insertion of human chromosome 6, and probes were made from both types of cells and labeled with two different fluors. These probes were mixed and applied to a microarray of 1161 cDNAs selected to study tumor suppression and the differential expression values provided valuable information about the molecular pathology of this tumor. For example, elevated expression of the WAF1 (p21) gene, which mediates *p53* tumor suppression, was observed only from non-tumorigenic probe. Likewise, elevated levels of human brown locus protein gene were observed only for tumorigenic cells.

Complementary DNA microarrays containing virtually all of the genes of *S. cerevisiae* have been fabricated[14]. This permitted the genome-wide study of the effects of the diauxic shift from anaerobic to aerobic metabolism under glucose limitation and the concomitant switch to ethanol as a carbon source. The significance of this work is that it mapped the changes in expression of genes with known function to their metabolic pathways and vividly showed which metabolic pathways were reprogrammed by the shift. For example, there were large induction in the gene encoding alcohol dehydrogenase and a corresponding shutdown of acetaldehyde dehydrogenase as the yeast cell turned to ethanol as the carbon source. Additionally the expression patterns of many previously unknown genes were obtained.

Shalon et al. produced arrays of previously mapped *S cerevisiae* genomic DNA fragments and hybridized these to chromosomal probes made from either the six largest chromosomes or the ten smallest chromosomes, labeled with lissamine or fluorescein, respectively[4]. Karyotypes of the sixteen chromosomes were produced in which the color of each segment, red or green, indicated whether the target DNA mapped to the six largest or ten smallest chromosomes, respectively. Ninety-five percent of the arrayed clones corresponded to previously published mapped positions.

Drmanac et al. have used comparatively massive cDNA microarrays for large-scale gene discovery in infant brain tissue[9]. In contrast to the work of Schena et al., probes were applied one at a time rather than in a mixture. The clones were statistically sorted into clusters each containing an expressed gene. 73,536 cDNA clones from infant brain libraries were exposed to 200–320 probes and the clusters were analyzed by the number of probes scored. 19,726 genes were identified and the data indicated that a further 20,000 may be expressed at low levels. These protocols are used for large-scale, commercial, gene screening in which microarrays of 55,000 cDNAs screen 800,000 clones per month (R. Drmanac, personal communication).

Similarly Milosavljevic et al. demonstrated genome-wide sequence recognition in the *Escherichia coli* genome via genomic DNA arrays[11]. Nine hundred and ninety-seven short oligonucleotide probes were hybridized with 15,328 randomly selected genomic clones. Lists were compiled of the probes that scored with each clone and were compared with a database of *E. coli* sequence data. 14.6 Mb of sequence structure was recognized in one experiment.

Use of oligonucleotide arrays for gene discovery and expression. In a direct parallel to the work of DeRisi et al. for the study of the whole genome of *S. cerevisiae* using cDNA microarrays[14], oligonucleotide arrays have been used for the same purpose by Wodicka et al.[22] Four GeneChips were used with a total of 260,000 25-mer oligonucleotide probes which covered every open reading frame (ORF) of the yeast genome. Each chip supported 65,000 probe sites (features). For greater accuracy each feature was synthesized with a neighbor that was a closely related mismatch differing by one central base. The signal from the mismatch probe was subtracted from the perfect match probe and compensated for nonspe-

cific binding and background fluorescence. Yeast cells were grown n rich or minimal media. Ninety percent of the genes were expressed under both growth conditions including structural proteins and ribosomal proteins. Good sensitivity and linearity were reported resulting in a linear range of 0.05 to 6 copies per cell. Thirty-six mRNA were more abundant in the rich medium and 140 mRNA were more abundant in minimal medium. In addition to genes of known function, previously uncharacterized genes were detected.

The abundance of data generated by these experiments, in particular the number of genes characterized demonstrates that this chip protocol can be used for efficient resequencing of complex samples containing ORFs up to 1 kb in length. It remains to be seen whether or not the observed changes in expression levels between the two growth media can be correlated to metabolic pathways as was done for the diauxic shift of *S. cerevisiae*[9].

Previously, Lockhart et al. used oligonucleotide arrays to measure the expression levels of all cytokine mRNA in murine T cells challenged with the stimulant 4-phorbol-12-myristate-13-acetate[10]. High levels of γ-interferon plus lower levels of cytokines were observed and as predicted β-actin and glyceraldehyde-3-phosphate dehydrogenase levels did not change significantly. In separate calibration experiments the dynamic range was determined to be from 1:300,000 to 1:300 w/w total mRNA. The limit of detection is comparable to the 1:500,000 w/w measured with the cDNA arrays of Schena et al.[11]

Oligonucleotide arrays have been used for measurement of expression levels of bacterial genes[12]. An important aspect of this work was that bacterial mRNA could be expressed without purification of the total RNA. Reproducible purification of bacterial mRNA is very difficult because of its low concentration (4%) in total RNA. An array of 64,000 oligonucleotides probes was fabricated for the detection of 100 *Streptococcus pneumoniae* genes. Sensitivity was better than obtained by Northern blotting and the error between duplicate samples was 25% or less. The arrays were used to demonstrate induction of the competency genes *cinA*, *recA* and *lytA* at 30-, 18-, and 10-fold levels, respectively. Similarly, differential expression of *S. pneumoniae* in exponential and stationary phases was measured. In the stationary phase genes encoding for enzymes of polysaccharide capsule biosynthesis, long-chain fatty acid biosynthesis and cell division were expressed at significantly lower levels, as expected and four genes were induced. This demonstrated the detection of sets of coregulated bacterial genes by use of a single array and without purification of mRNA.

Detection of mutations and polymorphisms. Hacia et al. used a GeneChip containing 96,600 oligonucleotide probes to detect all possible heterozygous mutations in the 3.45 kb exon 11 of the *BRCA1* breast and ovarian cancer gene. Controls and samples were differentiated by the two color fluorescence protocol. Fifteen patient samples were analyzed with one false negative and eight single base pair polymorphisms were detected. These exciting results have to be viewed in the context of further work as the *BRCA1* gene exhibits mutations in 22 coding exons. However, as they only cover 5592 bp, it is likely that a suitable chip could be fabricated for detection of all the mutations as chip densities of 400,000 probes have been achieved on GeneChips[6].

Schoemaker et al.[7] used GeneChip technology to demonstrate the utility of 20 base tags, each unique to one mutant, for simultaneous detection of eleven *S. cerevisiae* mutant strains grown under a variety of conditions. Under conditions in which adenine was not present in the growth medium adenine mutant strains became progressively weaker relative to the other non-compromised strains and this was shown by weaker signals on the microarray. Similarly, growth in media without tryptophan resulted in no growth by tryptophan mutants. Now that the *S. cerevisiae* genome has been sequenced, the biological functions of its genome needs to be further studied, for example by creating suitable deletion strains and testing under a wide range of selection conditions. Molecular bar coding could greatly facilitate this process.

Lipshutz et al.[8] demonstrated that the GeneChip could be used for screening of mutations in the reverse transcriptase and protease genes in the HIV-1 virus. Such mutations can cause resistance to antibiotics such as AZT. Kozal et al.[9] used a GeneChip to study the occurrence of polymorphisms in the HIV-1 clade protease gene in patients who had not been exposed to protease inhibitors. GeneChip results were checked with those obtained by Sanger sequencing. One hundred and fourteen samples were analyzed and the agreement between the two methods was excellent, 98%. A large degree of polymorphism was observed.

Mutations in the cystic fibrosis transmembrane conductance regulator (CFTR) were studied using the Affymetrix chip[6]. An array was designed to detect known deletion, insertion or base substitution mutations in exons 10 and 11 of CFTR. Ten unknown patient samples were tested and the results were confirmed exactly by PCR product restriction fragment analysis performed by independent workers.

Chee et al.[10] fabricated a GeneChip containing 135,000 25-mer probes for the probing of the 16.6 kb human mitochondrial genome. Two color fluorescence was used to compare mutated mitochondrial DNA (mtDNA) to control mtDNA. Mitochondrial genomes from 10 individuals were analyzed and 505 polymorphisms were automatically detected. Each sample could be read in 12 minutes. It was estimated that during a working day 40 mtDNA genomes could be read compared to two by a modern gel sequencer.

β-thalassemia mutations in blood samples have been detected by oligonucleotide microarrays in which presynthesized decamers were immobilized on gel-coated glass plates[11]. Mutations at three positions in the β-globin gene were detected by a 10-mer microarray. Contiguous stacking hybridization was then used to enhance detection of the IVS-1-1 (G→A) mutation by addition of soluble pentamer probe.

Heller and coworkers used positive fields to increase the transport rate of negatively charged probe, which increased the hybridization rate 10-fold[12]. A sample oligonucleotide was directed to electrode C4 by application of a positive field and bound to surface streptavidin. A negative potential was subsequently applied to repel unbound sample and the chip was washed with cysteine buffer. The process was repeated at electrode C5 but with a sample oligonucleotide that contained a single base pair mismatch. Bodipy Texas Red-conjugated probe was then applied under a positive potential, hybridized and unbound probe was removed by washing. The electodes were covered with electrolyte, a negative potential was applied and then a pulsed current was applied, in order to dissociate duplexes containing mutants. Single base pair mismatches were detected in less than 15 seconds by fluorescence detection[13].

Mapping genomic libraries. GeneChips have been used for mapping genomic libraries by determining the order of overlapping clones[14]. *S. cerevisiae* cosmid DNA was prepared from twelve genomic clones and a restriction enzyme was used to capture tetramer markers at EarI sites. After PCR amplification, labeling with fluorescent marker and production of ssDNA, the product was hybridized to a 256 feature array. Fluorescence intensities were normalized and a correlation score determined for each adjacent clone pair by statistical analysis. The ten cosmids that gave the strongest signals were arranged as a continuous sequence, and in the correct order, by using correlation scores in a simulated annealing procedure. This demonstrated the utility of this technique to map clones in a highly parallel fashion. Because all the chemical reactions for each clone were implemented in a single test tube, rapid parallel

assays of many clones could be envisioned. Furthermore, all the assay steps were amenable to automation for increased throughput. The authors estimated that a single operator could map several hundred clones per day.

Conclusion

DNA chip technology is rapidly advancing and applications to diagnostics (mutation detection), gene discovery, gene expression and mapping have been convincingly demonstrated. Format I arrays using bound DNA have been used for large-scale screening and expression studies. Format II, oligonucleotide arrays have been shown to be useful for rapid detection of mutations in *BRCA1*, HIV-1, cystic fibrosis and β-thalassemia genes and expression monitoring, gene discovery and mapping. The application of electric fields to increase hybridization rate and subsequently to denature duplexes containing single base pair mismatches has been demonstrated.

U.S. patents available on-line at http://www.patent.womplex.ibm.com

1. Southern, E.M. 1996. DNA chips: analysing sequence by hybridization to oligonucleotides on a large scale. *Trends Genet.* 12(3):110–115.
2. O'Donnell-Maloney, M.J., Smith, C.L., and Cantor, C.C. 1996. The development of microfabricated arrays for DNA sequencing and analysis. *Trends Biotech.* 14:401–407.
3. Ginot, F. 1997. Oligonucleotide microarrays for identification of unknown mutations: how far from reality? *Human Mutation* 10:1–10.
4. Drmanac, R., Drmanac, S., Labat, I., Crkvenjakov, R., Vicentric, A., and Gemmell, A. 1992. Sequencing by hybridization: towards an automated sequencing of one million M13 clones arrayed on membranes. *Electrophoresis* 13:566–573.
5. Schena, M., Shalon, D., Davis, R.W., and Brown, P.O. 1995. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270:467–470.
6. Wallraff, G., Labadie, J., Brock., P., DiPietro, R., Nguyen, T., Huynh, T., et al. 1997. DNA sequencing on a chip. *Chemtech* February, pp.22–32.
7. Schena, M. 1998. Genome analysis with gene expression microarrays. *Bioessays* 18:427–431.
8. Shalon, D., Smith, J.S., and Brown, P.O. 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6:639–645.
9. Drmanac, S., Stavropoulos, N.A., Labat, I., Vonau, J., Hauser, B., Soares, M.B., and Drmanac, R. 1996. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. *Genomics* 37:29–40.
10. Drmanac, S. and Drmanac, R. 1994. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. *BioTechniques* 17:328–336.
11. Milosavljevic, A., Savkovic, S., Crkvenjakov, R., Salbego, D., Serrato, H., Kreuzer, H., et al. 1996. DNA sequence recognition by hybridization to short oligomers: experimental verification of the method on the *E. coli* genome. *Genomics* 37:77–86.
12. Fodor, S.P.A., Read, L.J., Pirrung, M.C., Stryer, L., Lu, A.M., and Solas, D..1991. Light-directed, spatially addressable parallel chemical synthesis. *Science* 251:767–773.
13. Pease, A.N., Solas, D., Sullivan, E.J., Cronin, M.T., Holmes, C.P., and Fodor, S.P.A. 1994. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc. Natl. Acad. Sci. USA* 91:5022–5026.
14. Hacia, G.H., Brody, L.C., Chee, M.S., Fodor, S.P.A., and Collins, F.S. 1996. Detection of heterozygous mutations in BRCA1 using high-density oligonucleotide arrays and two color fluorescence analysis. *Nat. Genet.* 14:441–447.
15. McGall, G., Labadie, J., Brook, P., Wallraff, G., Nguyen, T., and Hinsberg, W. 1996. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc. Nat. Acad. Sci. USA* 93:13555–13560.
16. Yershov, K., Barsky, V., Belgovskiy, A., Kirillov, E., Kreindlin, E., Ivanov, I., et al. 1996. DNA analysis and diagnostics on oligonucleotide chips. *Genetics* 93:4913–4918.
17. Heller, M.J. and Tu, E. 1993. Active programmable electronic devices for molecular biomolecules. US Pat. 5605662.
18. DeRisi, J.L., Iyer, V.R., and Brown, P.O. 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 270:680–686.
19. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P.O., and Davis, R.O. 1996. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc .Nat. Acad. Sci. USA* 93:10614–10619.
20. Heller, R.N., Schena, M., Chai, A., Shalon, D., Bedilion, T., Gilmore, J., et al. 1997. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94:2150–2155.
21. DeRisi, J., Penland, L., Brown, P.O., Bittner, M.L., Meltzer, P.S., Ray, M., et al. 1996. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat. Genet.* 14:457–460.
22. Iyer, V., DeRisi, J., Eisen, M., Ross, D., Spellman, P., Hudson, J.Jr., et al. 1997. Use of DNA microarrays to monitor differential gene expression in yeast and humans. *Fed. Am. Soc. Exp. Bio.* 11:1126.
23. Wodicka, L., Dong, H., Mittmann, M., Ho, M-H., and Lockhart, D.J. 1997. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnology* 15:1–15.
24. Lockhart, D.J., Dong, H., Byrne, M.C., Follettie, M.T., Gallo, M.V., Chee, M.S., et al. 1996. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnology* 14:1675–1680.
25. De Saizieu, A., Certa, U., Warrington, J., Gray, C., Keck, W., and Mous, J. 1998. Bacterial transcript imaging by hybridization of total RNA to oligonucleotide arrays. *Nature Biotechnology* 16:45–48.
26. Stipp, D. 1997. Gene chip breakthrough. *Fortune* 3/31:56–73.
27. Shoemaker, D.D., Lashkari, D.A., Morris, D., Mittmann, M., and Davis, R.W. 1996. Quantitative phenotypic analysis of yeast deletion mutants using a highly parallel molecular bar-coding strategy. *Nat. Genet.* 14:450–456.
28. Lipshutz, R.J., Morris, D., Chee, M., Hubbell, E., Kozal, M.J., Shah, N., et al. 1995. Using oligonucleotide probe arrays to access genetic diversity. *Biotechniques* 19:442–447.
29. Kozal, M.J., Shah, N., Shen, N., Yang, R., Fucini, R., Merigan, T.C., et al. 1996. Extensive polymorphism observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. *Nature Medicine* 2:753–759.
30. Cronin, M.T., Fucini, R.V., Kim, S.M., Masino, R.S., Wespi, R.M., and Miyada, C.G. 1996. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. *Human Mut.* 7:244–255.
31. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., et al. 1996. Accessing genetic information with high-density DNA arrays. *Science* 274:610–613.
32. McIntyre, P.E. 1996. Microfabrication technology for DNA sequencing. *Trends Biotech.* 14:59–73.
33. Sosnowski, R.G., Tu, E., Butler, W.F., O'Connell, J.P., and Heller, M.J. 1997. Rapid determination of single base mismatch mutations in DNA hybrids by direct electric field control. *Proc. Nat. Acad. Sci. USA* 94:1119–1123.
34. Sapolsky, R.J. and Lipshutz, R.J. 1996. Mapping genomic library clones using oligonucleotide arrays. *Genomics* 33:445–456.

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for U49729

GI	Version	Update Date	Status	I	II
3320116	1	May 1 2000 2:49 PM	Live	◉	○
3320116	1	Jul 15 1998 11:18 AM	Dead	○	◉
1236563	n/a	Mar 27 1996 12:25 AM	Dead	○	○

Accession U49729.1 was first seen at NCBI on Mar 27 1996 12:25 AM

NCBI

Sequence Revision History

Revision history for X17163

GI	Version	Update Date	Status	I	II
57819	1	Apr 19 2005 3:31 AM	Live	◉	○
57819	1	Jul 24 2003 4:59 PM	Dead	○	◉
57819	1	Oct 18 2002 4:09 AM	Dead	○	○
57819	1	Mar 9 1999 7:46 AM	Dead	○	○
57819	1	May 26 1996 9:12 PM	Dead	○	○
57819	1	May 22 1995 5:19 PM	Dead	○	○
57819	1	Dec 1 1994 7:02 AM	Dead	○	○
57819	1	Sep 1 1993 2:23 PM	Dead	○	○
57819	1	Apr 21 1993 4:14 PM	Dead	○	○

Accession X17163.1 was first seen at NCBI on Apr 21 1993 4:14 PM

NCBI

Sequence Revision History

Revision history for L14680

GI	Version	Update Date	Status	I	II
408946	1	Nov 28 1994 11:12 PM	Live	◉	○
408946	1	Oct 4 1994 4:37 AM	Dead	○	◉
408946	1	Oct 19 1993 12:08 AM	Dead	○	○
405552	n/a	Oct 3 1993 12:12 AM	Dead	○	○

Accession L14680.1 was first seen at NCBI on Oct 3 1993 12:12 AM

NCBI

Sequence Revision History

Revision history for Y00396

GI	Version	Update Date	Status	I	II
55967	1	Apr 19 2005 6:35 AM	Live	◉	○
55967	1	Jul 24 2003 5:02 PM	Dead	○	◉
55967	1	Oct 18 2002 4:13 AM	Dead	○	○
55967	1	Jul 2 1999 12:16 AM	Dead	○	○
55967	1	Mar 30 1999 11:21 PM	Dead	○	○
55967	1	May 26 1996 8:31 PM	Dead	○	○
55967	1	May 22 1995 4:18 PM	Dead	○	○
55967	1	Dec 1 1994 6:16 AM	Dead	○	○
55967	1	Jun 16 1993 5:56 PM	Dead	○	○
55967	1	Apr 21 1993 3:27 PM	Dead	○	○

Accession Y00396.1 was first seen at NCBI on Apr 21 1993 3:27 PM

NCBI

Sequence Revision History

Show difference between I and II as GenBank/GenPept

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for X87257

GI	Version	Update Date	Status	I	II
836634	1	Apr 19 2005 6:04 AM	Live	◉	○
836634	1	Aug 5 2003 7:33 AM	Dead	○	◉
836634	1	Oct 14 2002 5:32 PM	Dead	○	○
836634	1	May 25 2000 1:00 PM	Dead	○	○
836634	1	Mar 9 1999 4:59 AM	Dead	○	○
836634	1	Oct 25 1996 2:48 AM	Dead	○	○
836634	1	May 26 1996 5:48 PM	Dead	○	○
836634	1	Jul 19 1995 12:32 AM	Dead	○	○
836634	1	Jun 8 1995 1:38 AM	Dead	○	○
836634	1	May 29 1995 1:04 AM	Dead	○	○

Accession X87257.1 was first seen at NCBI on May 29 1995 1:04 AM

NCBI

Sequence Revision History

Revision history for L20085

GI	Version	Update Date	Status	I	II
407036	1	Oct 4 1994 4:46 AM	Live	◉	○
407036	1	Oct 12 1993 12:03 AM	Dead	○	◉

Accession L20085.1 was first seen at NCBI on Oct 12 1993 12:03 AM

NCBI

Sequence Revision History

Revision history for M26744

GI	Version	Update Date	Status	I	II
204915	1	Oct 2 1995 10:50 AM	Live	◉	○
204915	1	Oct 4 1994 5:17 AM	Dead	○	◉
204915	1	Apr 27 1993 9:08 PM	Dead	○	○

Accession M26744.1 was first seen at NCBI on Apr 27 1993 9:08 PM

NCBI

Sequence Revision History

Revision history for X00469

GI	Version	Update Date	Status	I	II
56043	1	Apr 19 2005 1:54 AM	Live	◉	○
56043	1	Jul 24 2003 4:56 PM	Dead	○	◉
56043	1	Oct 18 2002 4:02 AM	Dead	○	○
56043	1	Mar 9 1999 7:26 AM	Dead	○	○
56043	1	May 26 1996 8:35 PM	Dead	○	○
56043	1	May 22 1995 4:21 PM	Dead	○	○
56043	1	Dec 1 1994 6:18 AM	Dead	○	○
56043	1	Sep 1 1993 12:04 PM	Dead	○	○
56043	1	Apr 21 1993 3:29 PM	Dead	○	○

Accession X00469.1 was first seen at NCBI on Apr 21 1993 3:29 PM

NCBI

Sequence Revision History

Revision history for U09540

GI	Version	Update Date	Status
1039376	1	Oct 25 1995 1:08 AM	Live

Accession U09540.1 was first seen at NCBI on Oct 25 1995 1:08 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for M34452

GI	Version	Update Date	Status	I	II
203679	1	Oct 4 1994 4:49 AM	Live	◉	○
203679	1	Apr 27 1993 8:36 PM	Dead	○	◉

Accession M34452.1 was first seen at NCBI on Apr 27 1993 8:36 PM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for M10161

GI	Version	Update Date	Status	I	II
203777	1	Oct 4 1994 4:51 AM	Live	◉	○
203777	1	Apr 27 1993 8:38 PM	Dead	○	◉

Accession M10161.1 was first seen at NCBI on Apr 27 1993 8:38 PM

NCBI

Sequence Revision History

Revision history for X07259

GI	Version	Update Date	Status	I	II
56046	1	Apr 19 2005 2:32 AM	Live	◉	○
56046	1	Jul 24 2003 5:15 PM	Dead	○	◉
56046	1	Oct 18 2002 4:31 AM	Dead	○	○
56046	1	Mar 9 1999 7:26 AM	Dead	○	○
56046	1	May 26 1996 8:35 PM	Dead	○	○
56046	1	May 22 1995 4:21 PM	Dead	○	○
56046	1	Dec 1 1994 6:18 AM	Dead	○	○
56046	1	Sep 1 1993 1:25 PM	Dead	○	○
56046	1	Apr 21 1993 3:29 PM	Dead	○	○

Accession X07259.1 was first seen at NCBI on Apr 21 1993 3:29 PM

NCBI

Sequence Revision History

Revision history for K03249

GI	Version	Update Date	Status	I	II
206068	1	Oct 4 1994 5:42 AM	Live	◉	○
206068	1	Apr 27 1993 9:39 PM	Dead	○	◉

Accession K03249.1 was first seen at NCBI on Apr 27 1993 9:39 PM

NCBI

Sequence Revision History

Revision history for M88592

GI	Version	Update Date	Status	I	II
206317	1	Oct 4 1994 5:48 AM	Live	◉	○
206317	1	Apr 27 1993 9:46 PM	Dead	○	◉

Accession M88592.1 was first seen at NCBI on Apr 27 1993 9:46 PM

NCBI

Sequence Revision History

Revision history for J02752

GI	Version	Update Date	Status	I	II
202677	1	Oct 4 1994 4:27 AM	Live	◉	○
202677	1	Apr 27 1993 8:09 PM	Dead	○	◉

Accession J02752.1 was first seen at NCBI on Apr 27 1993 8:09 PM

NCBI

Sequence Revision History

Revision history for U58829

GI	Version	Update Date	Status	I	II
1435202	1	Jun 7 2002 4:50 PM	Live	◉	○
1435202	1	Jan 7 1997 12:10 AM	Dead	○	◉
1435202	1	Jul 19 1996 1:53 AM	Dead	○	○

Accession U58829.1 was first seen at NCBI on Jul 19 1996 1:53 AM

NCBI

Sequence Revision History

Revision history for X15096

GI	Version	Update Date	Status	I	II
57707	1	Sep 9 2004 11:26 PM	Live	◉	○
57707	1	Jul 24 2003 5:02 PM	Dead	○	◉
57707	1	Oct 18 2002 4:16 AM	Dead	○	○
57707	1	Mar 9 1999 7:45 AM	Dead	○	○
57707	1	May 26 1996 9:10 PM	Dead	○	○
57707	1	May 22 1995 5:16 PM	Dead	○	○
57707	1	Dec 1 1994 7:00 AM	Dead	○	○
57707	1	Sep 1 1993 2:01 PM	Dead	○	○
57707	1	Apr 21 1993 4:11 PM	Dead	○	○

Accession X15096.1 was first seen at NCBI on Apr 21 1993 4:11 PM

NCBI

Sequence Revision History

Revision history for U30186

GI	Version	Update Date	Status
915541	1	Aug 1 1995 12:17 AM	Live

Accession U30186.1 was first seen at NCBI on Aug 1 1995 12:17 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for L32591

GI	Version	Update Date	Status	I	II
799327	1	Apr 8 1996 3:22 PM	Live	◉	○
799327	1	Apr 8 1996 3:19 PM	Dead	○	◉
799327	1	May 26 1995 7:20 PM	Dead	○	○
799327	1	May 8 1995 3:19 PM	Dead	○	○
762838	n/a	Apr 7 1995 12:49 AM	Dead	○	○
483413	n/a	Oct 4 1994 5:01 AM	Dead	○	○
483413	n/a	May 8 1994 1:54 AM	Dead	○	○

Accession L32591.1 was first seen at NCBI on May 8 1994 1:54 AM

NCBI

Sequence Revision History

Revision history for M76704

GI	Version	Update Date	Status	I	II
206684	1	Oct 4 1994 5:56 AM	Live	◉	○
206684	1	Apr 27 1993 9:55 PM	Dead	○	◉

Accession M76704.1 was first seen at NCBI on Apr 27 1993 9:55 PM

NCBI

Sequence Revision History

Revision history for K01931

GI	Version	Update Date	Status	I	II
204494	1	Oct 4 1994 5:07 AM	Live	◉	○
204494	1	Apr 27 1993 8:57 PM	Dead	○	◉

Accession K01931.1 was first seen at NCBI on Apr 27 1993 8:57 PM

NCBI

Sequence Revision History

Revision history for X67654

GI	Version	Update Date	Status	I	II
56264	1	Aug 27 2003 1:16 PM	Live	◉	○
56264	1	Mar 9 1999 7:27 AM	Dead	○	◉
56264	1	May 27 1996 1:42 PM	Dead	○	○
56264	1	May 22 1995 4:27 PM	Dead	○	○
56264	1	Dec 1 1994 6:24 AM	Dead	○	○
56264	1	Jul 27 1993 11:32 PM	Dead	○	○
56264	1	Apr 21 1993 3:35 PM	Dead	○	○

Accession X67654.1 was first seen at NCBI on Apr 21 1993 3:35 PM

NCBI

Sequence Revision History

Revision history for U73174

GI	Version	Update Date	Status
1657631	1	Nov 4 1996 12:08 AM	Live

Accession U73174.1 was first seen at NCBI on Nov 4 1996 12:08 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for J05405

GI	Version	Update Date	Status	I	II
204626	1	Oct 4 1994 5:11 AM	Live	◉	○
204626	1	Apr 27 1993 9:01 PM	Dead	○	◉

Accession J05405.1 was first seen at NCBI on Apr 27 1993 9:01 PM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for L16764

GI	Version	Update Date	Status	I	II
294567	1	Apr 6 1994 11:36 AM	Live	◉	○
294567	1	Jun 12 1993 3:14 AM	Dead	○	◉

Accession L16764.1 was first seen at NCBI on Jun 12 1993 3:14 AM

NCBI

Sequence Revision History

Revision history for Y00497

GI	Version	Update Date	Status	I	II
56690	1	Apr 19 2005 6:36 AM	Live	◉	○
56690	1	Jul 24 2003 5:08 PM	Dead	○	◉
56690	1	Oct 18 2002 4:20 AM	Dead	○	○
56690	1	Mar 9 1999 7:30 AM	Dead	○	○
56690	1	May 27 1996 1:54 PM	Dead	○	○
56690	1	Apr 1 1995 12:01 AM	Dead	○	○
56690	1	Mar 31 1995 3:02 AM	Dead	○	○
56690	1	Dec 1 1994 6:34 AM	Dead	○	○
56690	1	Sep 1 1993 5:25 PM	Dead	○	○
56690	1	Apr 21 1993 3:45 PM	Dead	○	○

Accession Y00497.1 was first seen at NCBI on Apr 21 1993 3:45 PM

NCBI

Sequence Revision History

Revision history for M16975

GI	Version	Update Date	Status	I	II
204472	1	Oct 4 1994 5:07 AM	Live	◉	○
204472	1	Apr 27 1993 8:57 PM	Dead	○	◉

Accession M16975.1 was first seen at NCBI on Apr 27 1993 8:57 PM

NCBI

Sequence Revision History

Revision history for M27315

GI	Version	Update Date	Status	I	II
343181	1	Mar 5 2001 3:56 PM	Live	◉	○
343181	1	May 29 1996 12:17 PM	Dead	○	◉
343181	1	May 24 1996 1:33 AM	Dead	○	○
343181	1	May 24 1995 2:13 AM	Dead	○	○
343181	1	Apr 28 1995 12:45 AM	Dead	○	○
343181	1	Oct 3 1994 8:22 AM	Dead	○	○
343181	1	Aug 4 1993 1:26 AM	Dead	○	○

Accession M27315.1 was first seen at NCBI on Aug 4 1993 1:26 AM

NCBI

Sequence Revision History

Revision history for D90102

This ID replaces sequence(s) **1) M32987 (See Rev. history)**
Common Rev. history

GI	Version	Update Date	Status	I	II
220766	1	Dec 17 2002 12:05 AM	Live	◉	○
220766	1	Jul 24 2002 10:59 AM	Dead	○	○
220766	1	Feb 2 2000 7:38 PM	Dead	○	○
220766	1	Mar 17 1999 10:06 PM	Dead	○	○
220766	1	Jun 5 1997 3:07 PM	Dead	○	○
220766	1	Sep 2 1993 3:27 AM	Dead	○	○
220766	1	Apr 29 1993 10:54 AM	Dead	○	○

Accession D90102.1 was first seen at NCBI on Apr 29 1993 10:54 AM

Revision history for D90102

GI	Version	Update Date	Status	I	II
25293660	n/a	Mar 10 2005 12:41 PM	Live	○	◉
25293660	n/a	Nov 25 2002 2:29 PM	Dead	○	○

Accession D90102 was first seen at NCBI on Nov 25 2002 2:29 PM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for X52625

GI	Version	Update Date	Status	I	II
55946	1	Apr 19 2005 3:41 AM	Live	◉	○
55946	1	Jul 24 2003 4:48 PM	Dead	○	◉
55946	1	Jun 18 2003 11:12 PM	Dead	○	○
55946	1	Oct 18 2002 3:51 AM	Dead	○	○
55946	1	Mar 9 1999 7:25 AM	Dead	○	○
55946	1	May 26 1996 8:30 PM	Dead	○	○
55946	1	May 22 1995 4:17 PM	Dead	○	○
55946	1	Dec 1 1994 6:15 AM	Dead	○	○
55946	1	Sep 1 1993 2:39 PM	Dead	○	○
55946	1	Apr 21 1993 3:27 PM	Dead	○	○

Accession X52625.1 was first seen at NCBI on Apr 21 1993 3:27 PM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for L27129

GI	Version	Update Date	Status	I	II
493212	1	Oct 4 1994 6:01 AM	Live	◉	○
493212	1	May 25 1994 3:47 AM	Dead	○	◉

Accession L27129.1 was first seen at NCBI on May 25 1994 3:47 AM

NCBI

Sequence Revision History

Revision history for U89282

GI	Version	Update Date	Status
1932816	1	Apr 11 1997 12:16 AM	Live

Accession U89282.1 was first seen at NCBI on Apr 11 1997 12:16 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for D14014

GI	Version	Update Date	Status	I	II
413908	1	Jan 9 2003 3:20 AM	Live	◉	○
413908	1	Jul 25 2002 2:10 PM	Dead	○	◉
413908	1	Mar 17 1999 10:05 PM	Dead	○	○
413908	1	Jun 5 1997 3:02 PM	Dead	○	○
413908	1	Dec 18 1993 12:24 AM	Dead	○	○
413908	1	Nov 4 1993 12:14 AM	Dead	○	○

Accession D14014.1 was first seen at NCBI on Nov 4 1993 12:14 AM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for L26267

GI	Version	Update Date	Status	I	II
425471	1	Aug 15 1994 4:55 PM	Live	◉	○
425471	1	May 28 1994 12:22 AM	Dead	○	◉
425471	1	Apr 23 1994 12:16 AM	Dead	○	○
425471	1	Nov 25 1993 1:54 AM	Dead	○	○

Accession L26267.1 was first seen at NCBI on Nov 25 1993 1:54 AM

NCBI

Sequence Revision History

Revision history for U73142

GI	Version	Update Date	Status	I	II
1621646	1	Nov 5 1998 8:14 AM	Live	◉	○
1621646	1	Oct 22 1996 12:14 AM	Dead	○	◉

Accession U73142.1 was first seen at NCBI on Oct 22 1996 12:14 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for M61177

GI	Version	Update Date	Status	I	II
204053	1	Mar 26 2002 1:52 PM	Live	◉	○
204053	1	Mar 9 1995 6:06 PM	Dead	○	◉
204053	1	Oct 4 1994 4:57 AM	Dead	○	○
204053	1	Apr 27 1993 8:46 PM	Dead	○	○

Accession M61177.1 was first seen at NCBI on Apr 27 1993 8:46 PM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for X12752

GI	Version	Update Date	Status	I	II
510928	1	Jul 24 2003 4:51 PM	Live	◉	○
510928	1	Oct 18 2002 3:54 AM	Dead	○	◉
510928	1	Mar 9 1999 7:25 AM	Dead	○	○
510928	1	May 26 1996 8:29 PM	Dead	○	○
510928	1	May 22 1995 4:17 PM	Dead	○	○
510928	1	Dec 1 1994 6:15 AM	Dead	○	○
510928	1	Aug 17 1994 2:11 AM	Dead	○	○
510928	1	Jul 26 1994 11:19 AM	Dead	○	○
510928	1	Jul 17 1994 12:35 AM	Dead	○	○
55925	n/a	Sep 1 1993 1:36 PM	Dead	○	○
55925	n/a	Apr 21 1993 3:26 PM	Dead	○	○

Accession X12752.1 was first seen at NCBI on Apr 21 1993 3:26 PM

NCBI

Sequence Revision History

Revision history for U66479

GI	Version	Update Date	Status	I	II
1710130	1	Sep 24 1998 2:44 PM	Live	◉	○
1710130	1	Dec 10 1996 12:19 AM	Dead	○	◉
1710130	1	Dec 9 1996 5:23 PM	Dead	○	○

Accession U66479.1 was first seen at NCBI on Dec 9 1996 5:23 PM

NCBI

Sequence Revision History

Revision history for J04791

GI	Version	Update Date	Status	I	II
205807	1	May 11 1995 12:54 AM	Live	◉	○
205807	1	Oct 4 1994 5:37 AM	Dead	○	◉
205807	1	Apr 27 1993 9:32 PM	Dead	○	○

Accession J04791.1 was first seen at NCBI on Apr 27 1993 9:32 PM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for X13058

GI	Version	Update Date	Status	I	II
56828	1	Apr 19 2005 2:38 AM	Live	◉	○
56828	1	Jul 24 2003 5:12 PM	Dead	○	◉
56828	1	Oct 18 2002 4:24 AM	Dead	○	○
56828	1	Mar 9 1999 7:32 AM	Dead	○	○
56828	1	May 26 1996 8:42 PM	Dead	○	○
56828	1	May 22 1995 4:41 PM	Dead	○	○
56828	1	Dec 1 1994 6:37 AM	Dead	○	○
56828	1	Sep 1 1993 1:40 PM	Dead	○	○
56828	1	Apr 21 1993 3:49 PM	Dead	○	○

Accession X13058.1 was first seen at NCBI on Apr 21 1993 3:49 PM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for Y00047

This ID replaces sequence(s) **1) M24604 (See Rev. history)**
Common Rev. history

GI	Version	Update Date	Status	I	II
56861	1	Apr 19 2005 6:36 AM	Live	◉	○
56861	1	Jul 24 2003 5:29 PM	Dead	○	◉
56861	1	May 9 2003 7:43 AM	Dead	○	○
56861	1	Mar 9 1999 7:32 AM	Dead	○	○
56861	1	Oct 4 1996 1:18 AM	Dead	○	○
56861	1	May 26 1996 8:43 PM	Dead	○	○
56861	1	May 22 1995 4:43 PM	Dead	○	○
56861	1	Dec 1 1994 6:38 AM	Dead	○	○
56861	1	Sep 1 1993 5:22 PM	Dead	○	○
56861	1	Apr 21 1993 3:50 PM	Dead	○	○

Accession Y00047.1 was first seen at NCBI on Apr 21 1993 3:50 PM

NCBI

Sequence Revision History

Revision history for M81855

GI	Version	Update Date	Status	I	II
205360	1	Oct 4 1994 5:27 AM	Live	◉	○
205360	1	Apr 27 1993 9:20 PM	Dead	○	◉

Accession M81855.1 was first seen at NCBI on Apr 27 1993 9:20 PM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

NCBI

Sequence Revision History

Revision history for D38380

GI	Version	Update Date	Status	I	II
1854475	1	Jul 25 2002 2:09 PM	Live	◉	○
1854475	1	Jun 7 1999 10:29 AM	Dead	○	◉
1854475	1	Mar 17 1999 10:08 PM	Dead	○	○
1854475	1	Apr 18 1997 12:47 AM	Dead	○	○
1854475	1	Mar 28 1997 12:40 AM	Dead	○	○
1854475	1	Feb 28 1997 12:35 AM	Dead	○	○
551297	n/a	May 9 1996 6:33 PM	Dead	○	○
551297	n/a	Mar 18 1996 12:19 AM	Dead	○	○
551297	n/a	Sep 29 1994 11:33 AM	Dead	○	○

Accession D38380.1 was first seen at NCBI on Sep 29 1994 11:33 AM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for V01222

This ID replaces sequence(s) **1) J00698 (See Rev. history)**
Common Rev. history

GI	Version	Update Date	Status	I	II
55627	1	Apr 19 2005 1:47 AM	Live	◉	○
55627	1	Mar 11 2005 9:29 AM	Dead	○	◉
55627	1	Mar 9 1999 7:23 AM	Dead	○	○
55627	1	Oct 4 1996 1:17 AM	Dead	○	○
55627	1	May 26 1996 8:22 PM	Dead	○	○
55627	1	Apr 8 1995 12:42 AM	Dead	○	○
55627	1	Dec 1 1994 6:08 AM	Dead	○	○
55627	1	Sep 1 1993 11:54 AM	Dead	○	○
55627	1	Apr 21 1993 3:19 PM	Dead	○	○

Accession V01222.1 was first seen at NCBI on Apr 21 1993 3:19 PM

NCBI

Sequence Revision History

Revision history for X69021

GI	Version	Update Date	Status	I	II
57254	1	Aug 27 2003 1:16 PM	Live	◉	○
57254	1	Mar 9 1999 7:35 AM	Dead	○	◉
57254	1	May 26 1996 8:56 PM	Dead	○	○
57254	1	May 22 1995 4:59 PM	Dead	○	○
57254	1	Dec 1 1994 6:48 AM	Dead	○	○
57254	1	Nov 18 1993 12:38 AM	Dead	○	○
57254	1	Sep 1 1993 5:11 PM	Dead	○	○
57254	1	Apr 21 1993 4:00 PM	Dead	○	○

Accession X69021.1 was first seen at NCBI on Apr 21 1993 4:00 PM

NCBI

Sequence Revision History

Revision history for X66539

GI	Version	Update Date	Status	I	II
395369	1	Aug 27 2003 1:15 PM	Live	◉	○
395369	1	Mar 9 1999 7:36 AM	Dead	○	◉
395369	1	Oct 28 1997 1:59 AM	Dead	○	○
395369	1	May 26 1996 9:00 PM	Dead	○	○
395369	1	May 22 1995 5:07 PM	Dead	○	○
395369	1	Dec 1 1994 6:51 AM	Dead	○	○
395369	1	Nov 18 1993 12:39 AM	Dead	○	○
395369	1	Sep 3 1993 3:21 PM	Dead	○	○

Accession X66539.1 was first seen at NCBI on Sep 3 1993 3:21 PM

NCBI

Sequence Revision History

Revision history for L09653

GI	Version	Update Date	Status	I	II
207289	1	Oct 4 1994 6:11 AM	Live	◉	○
207289	1	Jul 26 1993 9:31 PM	Dead	○	◉
207289	1	Apr 27 1993 10:12 PM	Dead	○	○

Accession L09653.1 was first seen at NCBI on Apr 27 1993 10:12 PM

NCBI

Sequence Revision History

Revision history for J05132

GI	Version	Update Date	Status	I	II
207602	1	Oct 4 1994 6:17 AM	Live	◉	○
207602	1	Apr 27 1993 10:20 PM	Dead	○	◉

Accession J05132.1 was first seen at NCBI on Apr 27 1993 10:20 PM

NCBI

Sequence Revision History

Revision history for D83796

GI	Version	Update Date	Status	I	II
1330262	1	Jul 25 2002 2:37 PM	Live	◉	○
1330262	1	Mar 17 1999 7:35 PM	Dead	○	○
1330262	1	Aug 14 1997 1:13 AM	Dead	○	○
1330262	1	Jun 5 1997 2:06 PM	Dead	○	○
1330262	1	Mar 17 1997 12:04 AM	Dead	○	○
1330262	1	May 24 1996 1:36 AM	Dead	○	○
1330262	1	May 23 1996 7:36 PM	Dead	○	○
1230552	n/a	Mar 19 1996 1:32 AM	Dead	○	○

Accession D83796.1 was first seen at NCBI on Mar 19 1996 1:32 AM

Revision history for D83796

GI	Version	Update Date	Status	I	II
25300874	n/a	Mar 10 2005 1:42 PM	Live	○	◉
25300874	n/a	Nov 25 2002 4:19 PM	Dead	○	○

Accession D83796 was first seen at NCBI on Nov 25 2002 4:19 PM

NCBI

Sequence Revision History

Revision history for J02635

GI	Version	Update Date	Status	I	II
202591	1	Oct 4 1994 4:25 AM	Live	◉	○
202591	1	Apr 27 1993 8:07 PM	Dead	○	◉

Accession J02635.1 was first seen at NCBI on Apr 27 1993 8:07 PM

NCBI

Sequence Revision History

Revision history for V01227

GI	Version	Update Date	Status	I	II
55776	1	Apr 19 2005 1:48 AM	Live	◉	○
55776	1	Jul 24 2003 5:20 PM	Dead	○	◉
55776	1	Oct 18 2002 4:41 AM	Dead	○	○
55776	1	Mar 9 1999 7:24 AM	Dead	○	○
55776	1	May 26 1996 8:25 PM	Dead	○	○
55776	1	May 22 1995 4:12 PM	Dead	○	○
55776	1	Dec 1 1994 6:11 AM	Dead	○	○
55776	1	Sep 1 1993 11:54 AM	Dead	○	○
55776	1	Apr 21 1993 3:22 PM	Dead	○	○

Accession V01227.1 was first seen at NCBI on Apr 21 1993 3:22 PM

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for X59051

GI	Version	Update Date	Status	I	II
57132	1	Sep 9 2004 11:30 PM	Live	◉	○
57132	1	Jul 24 2003 5:31 PM	Dead	○	◉
57132	1	Oct 18 2002 4:54 AM	Dead	○	○
57132	1	Mar 9 1999 7:34 AM	Dead	○	○
57132	1	Aug 6 1996 12:32 AM	Dead	○	○
57132	1	May 26 1996 8:52 PM	Dead	○	○
57132	1	May 22 1995 4:51 PM	Dead	○	○
57132	1	Dec 1 1994 6:45 AM	Dead	○	○
57132	1	Sep 1 1993 3:38 PM	Dead	○	○
57132	1	Apr 21 1993 3:56 PM	Dead	○	○

Accession X59051.1 was first seen at NCBI on Apr 21 1993 3:56 PM

NCBI

Sequence Revision History

Revision history for X68199

GI	Version	Update Date	Status	I	II
56732	1	Apr 19 2005 5:17 AM	Live	◉	○
56732	1	Jul 24 2003 5:07 PM	Dead	○	◉
56732	1	Oct 18 2002 4:20 AM	Dead	○	○
56732	1	Mar 9 1999 7:30 AM	Dead	○	○
56732	1	Dec 10 1997 2:33 AM	Dead	○	○
56732	1	May 26 1996 8:39 PM	Dead	○	○
56732	1	May 22 1995 4:39 PM	Dead	○	○
56732	1	Dec 1 1994 6:35 AM	Dead	○	○
56732	1	Sep 1 1993 5:07 PM	Dead	○	○
56732	1	Jun 16 1993 1:40 PM	Dead	○	○
56732	1	Apr 21 1993 3:46 PM	Dead	○	○

Accession X68199.1 was first seen at NCBI on Apr 21 1993 3:46 PM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

NCBI

Sequence Revision History

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Revision history for M86443

GI	Version	Update Date	Status	I	II
204660	1	Oct 4 1994 5:12 AM	Live	◉	○
204660	1	Apr 27 1993 9:01 PM	Dead	○	◉

Accession M86443.1 was first seen at NCBI on Apr 27 1993 9:01 PM

NCBI

Sequence Revision History

Revision history for D16554

GI	Version	Update Date	Status	I	II
471155	1	Jan 9 2003 3:17 AM	Live	◉	○
471155	1	Jul 25 2002 2:08 PM	Dead	○	◉
471155	1	Feb 2 2000 7:27 PM	Dead	○	○
471155	1	Mar 17 1999 10:08 PM	Dead	○	○
471155	1	Jun 5 1997 3:14 PM	Dead	○	○
471155	1	Feb 16 1995 12:21 AM	Dead	○	○
471155	1	Aug 2 1994 12:17 PM	Dead	○	○
471155	1	Apr 13 1994 1:00 AM	Dead	○	○

Accession D16554.1 was first seen at NCBI on Apr 13 1994 1:00 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for D00036

GI	Version	Update Date	Status	I	II
220874	1	Jul 25 2002 2:22 PM	Live	◉	○
220874	1	Mar 17 1999 10:08 PM	Dead	○	◉
220874	1	Jan 8 1998 3:32 AM	Dead	○	○
220874	1	Jun 5 1997 3:14 PM	Dead	○	○
220874	1	Apr 29 1993 10:58 AM	Dead	○	○

Accession D00036.1 was first seen at NCBI on Apr 29 1993 10:58 AM

Entrez

Search for Genes
LocusLink provides curated information for human, fruit fly, mouse, rat, and zebrafish

Help|FAQ

Batch Entrez: Upload a file of GI or accession numbers to retrieve protein or nucleotide sequences

Check sequence revision history

How to create WWW links to Entrez

LinkOut

Cubby

Related resources

BLAST

Reference sequence project

LocusLink

Clusters of orthologous groups

Protein reviews on the web

Disclaimer | Write to the Help Desk
NCBI | NLM|NIH

NCBI

Sequence Revision History

Revision history for X02231

GI	Version	Update Date	Status	I	II
56187	1	Apr 19 2005 1:57 AM	Live	◉	○
56187	1	Jul 24 2003 5:18 PM	Dead	○	◉
56187	1	Jun 12 2003 12:05 AM	Dead	○	○
56187	1	Oct 18 2002 4:36 AM	Dead	○	○
56187	1	Mar 9 1999 7:27 AM	Dead	○	○
56187	1	May 22 1995 4:25 PM	Dead	○	○
56187	1	Dec 17 1994 12:32 AM	Dead	○	○
56187	1	Dec 1 1994 6:22 AM	Dead	○	○
56187	1	Sep 1 1993 12:21 PM	Dead	○	○
56187	1	Apr 21 1993 3:32 PM	Dead	○	○

Accession X02231.1 was first seen at NCBI on Apr 21 1993 3:32 PM

NCBI

Sequence Revision History

Revision history for V01217

This ID replaces sequence(s) **1) J00691 (See Rev. history)**
Common Rev. history

GI	Version	Update Date	Status	I	II
55574	1	Apr 19 2005 1:47 AM	Live	◉	○
55574	1	Mar 22 2004 10:19 AM	Dead	○	◉
55574	1	Mar 9 1999 7:22 AM	Dead	○	○
55574	1	May 26 1996 8:20 PM	Dead	○	○
55574	1	May 22 1995 4:07 PM	Dead	○	○
55574	1	Dec 1 1994 6:06 AM	Dead	○	○
55574	1	Sep 1 1993 11:54 AM	Dead	○	○
55574	1	Jun 16 1993 4:52 PM	Dead	○	○
55574	1	Apr 21 1993 3:18 PM	Dead	○	○

Accession V01217.1 was first seen at NCBI on Apr 21 1993 3:18 PM

GBREL.TXT Genetic Sequence Data Bank
15 December 1997

NCBI-GenBank Flat File Release 104.0

Distribution CD-ROM Release Notes

1891953 loci, 1258290513 bases, from 1891953 reported sequences

This document describes the data written on GenBank flat file distribution CD-ROMs. If you have any questions or comments about GenBank, the CD-ROM, or this document, please contact NCBI via email at info@ncbi.nlm.nih.gov or:

GenBank
National Center for Biotechnology Information
National Library of Medicine, 38A, 8N805
8600 Rockville Pike
Bethesda, MD 20894
USA
Phone: (301) 496-2475
Fax: (301) 480-9241

==========

TABLE OF CONTENTS

==========

==========

1. INTRODUCTION

1.1 Release 104.0

The National Center for Biotechnology Information (NCBI) at the National Library of Medicine (NLM), National Institutes of Health (NIH) is responsible for producing and distributing the GenBank Sequence Database. NCBI handles all GenBank direct submissions and authors are advised to use the address below. Submitters are encouraged to use the free Sequin software package for sending sequence data, or the newly developed World Wide Web submission form. See Section 1.5 below for details.

****

The address for direct submissions to GenBank is:

GenBank Submissions
National Center for Biotechnology Information
Bldg 38A, Rm. 8N-803
8600 Rockville Pike
Bethesda, MD 20894

E-MAIL: gb-sub@ncbi.nlm.nih.gov

Updates and changes to existing GenBank records:

E-MAIL: update@ncbi.nlm.nih.gov

URL for the new GenBank submission tool - BankIt - on the World Wide Web:

http://www.ncbi.nlm.nih.gov/

(see Section 1.5 for additional details about submitting data to GenBank.)

****

GenBank Release 104.0 is a release of sequence data by NCBI in the GenBank flat file format. GenBank is a component of a tri-partite, international collaboration of sequence databases in the U.S., Europe, and Japan. The collaborating databases in Europe are the European Molecular Biology Laboratory (EMBL) at Hinxton Hall, UK, and the DNA Database of Japan (DDBJ) in Mishima, Japan. Sequence data is also incorporated from the Genome Sequence Data Base (GSDB), Santa Fe, NM. Patent sequences are incorporated through arrangements with the U.S. Patent and Trademark Office, and via the collaborating international databases from other international patent offices. The database is converted to various output formats, including the Flat File and Abstract Syntax Notation 1 (ASN.1) versions. The ASN.1 and Flat File forms of the data are also available by anonymous FTP to 'ncbi.nlm.nih.gov'.

1.2 Cutoff Date

This full release, 104.0, incorporates data available to the databases as of December 4, 1997. For more recent data, users are advised to download the update files by anonymous FTP to 'ncbi.nlm.nih.gov' or to search the updates via the e-mail server. For instructions on the use of the e-mail server, send mail message with the word 'help' in it to: retrieve@ncbi.nlm.nih.gov

1.3 Important Changes in Release 104.0

1.3.1 Organizational changes

Due to the growth in the number of EST sequences, the EST division is now being split into 19 pieces.

For the CD-ROM version of this release, the gbaut.idx "index" file has been arbitrarily split into two pieces because its size exceeds the capacity of a single disc. The file names of the pieces are gbaut1.idx and gbaut2.idx .

1.3.2 New PUBMED linetype

A new reference linetype (PUBMED) is legal as of GenBank Release 104.0. This linetype will be used for literature citations that have a PubMed database identifier but lack a MEDLINE Unique Identifier (MUID). Here is a mocked-up example of what such a reference might look like:

```
REFERENCE   1  (bases 1 to 1512)
  AUTHORS   Palus,J.A., Ludden,P.W. and Triplett,E.W.
  TITLE     Diazotrophic bacterial endophytes isolated from stems of Zea mays
            L. and Zea luxurians Iltis and Doebley
  JOURNAL   Plant Soil 186, 135-142 (1996)
  PUBMED    123456789
```

1.3.3 New source feature qualifiers

As of GenBank Release 104.0, two new qualifiers have been introduced for the source feature: /specimen_voucher and /focus .

Qualifier	/specimen_voucher="text"
Definition	an identifier of the individual or collection of the source organism and the place where it is currently stored, usually an institution.
Value format	"text"
Example	/specimen_voucher="Smith s. n. 4-IV-1995 (U. S. Natl. Herbarium)"

Qualifier	/focus
Definition	defines the preferred source feature for records that have more than one source feature
Value format	none
Example	/focus
Comment	this qualifier is to be used only if there is more than one source feature. The preferred source feature is used to determine which organism is displayed in the SOURCE and ORGANISM lines and to determine the GenBank division flatfile in which it is placed.

1.3.4 /translation and translation-related qualifiers

For historical reasons, some unusual usages of /translation and translation-related qualifiers have accumulated in the database. During the most recent DDBJ/EMBL/GenBank collaborative meeting, it was decided that:

/translation will be removed from all but the CDS feature

translation-related qualifiers (/codon, /codon_start, /transl_table, /transl_except, /exception) will no longer be allowed on non-CDS features.

The features involved in these changes include: exon, transit_peptide, sig_peptide, mat_peptide, C_region, D_segment, J_segment, N_region, S_region, V_region, and V_segment.

1.3.5 /pseudo on transit_peptide, mat_peptide, and mRNA features.

Starting with GenBank Release 104.0, the /pseudo qualifier is legal for the transit_peptide, mat_peptide, and mRNA features.

1.4 Upcoming Changes

1.4.1 Removal of 'index' files from CdRom distribution of Releases 105 and 106

To stay within the 12-disc limit of our CdRom production contract, the five 'index' files (gbacc.idx, gbaut.idx, gbgen.idx, gbjou.idx, and gbkey.idx) that accompany the sequence data files of GenBank releases will not be included in the CdRom distributions of Releases 105.0 and 106.0 . All five index files will be available by anonymous FTP from:

ftp://ncbi.nlm.nih.gov/genbank

1.4.2 GenBank CdRom distribution ends with Release 106.0

GenBank releases will no longer be distributed via CdRom after Release 106.0

in April of 1998; ftp will be the only method of distribution. All current CdRom subscribers have been informed of this change, which has been necessitated by the growth of the database.

1.4.3 Accession Number Format, NIDs, and PIDs

With GenBank Release 81.0 (February, 1994) NCBI introduced an integer identifier called a 'gi' for every sequence (DNA, RNA, protein translation) in the database. The purpose of this identifer is to track a sequence as it changes over time; a new gi is assigned to every sequence version, and pointers between old and new gis are established. gis originally appeared via the COMMENT and /note fields of the GenBank flatfile format.

When DDBJ and EMBL introduced similar sequence tracking methods, the more general terms 'nucleotide identifiers' (NIDs) and 'protein identifiers' (PIDs) were adopted, and new linetypes and qualifiers were defined for these types of identifiers.

NIDs and PIDs have drawbacks, however. They are large integer values that communicate no intrinsic meaning to database users. And they are really internal database keys not easily amenable to collaborative maintenance. For example, a protein translation issued a PID by DDBJ must still be assigned a 'gi' when received by NCBI, which leads to two PID /db_xref qualifiers on the corresponding CDS feature (one with a 'd' PID value and the other with a 'g' PID value).

For these reasons, DDBJ, EMBL, and GenBank have agreed to introduce a new system of identifiers for *both* nucleotide and protein sequences, of the form 'Accession.Version' (eg, AB000349.3). The accession portion of these identifiers is stable and will not change, but the version portion will be incremented whenever the underlying sequence changes.

Here is an example of how ACCESSION, NID, and /db_xref currently appear in a typical GenBank entry:

```
LOCUS       AAU36846      568 bp    DNA             PRI       26-OCT-1995
DEFINITION  Aotus azarai cytochrome c oxidase subunit II (COII) gene,
            mitochondrial gene encoding mitochondrial protein, partial cds.
ACCESSION   U36846
NID         g1040987
...
     CDS             <1..>568
                     /gene="COII"
                     /codon_start=1
                     /product="cytochrome c oxidase subunit II"
                     /db_xref="PID:g1040988"
```

During transition to the Accession.Version system, a new VERSION linetype and /protein_id qualifier will be introduced:

```
LOCUS       AAU36846      568 bp    DNA             PRI       26-OCT-1995
DEFINITION  Aotus azarai cytochrome c oxidase subunit II (COII) gene,
            mitochondrial gene encoding mitochondrial protein, partial cds.
ACCESSION   U36846
NID         g1040987
VERSION     U36846.1  GI:1040987
...
     CDS             <1..>568
                     /gene="COII"
                     /codon_start=1
```

```
                     /product="cytochrome c oxidase subunit II"
                     /protein_id = "AAA12345.1"
                     /db_xref="PID:g1040988"
                     /db_xref="GI:1040988"
```

And after the transition period is complete:

```
LOCUS       AAU36846      568 bp    DNA             PRI       26-OCT-1995
DEFINITION  Aotus azarai cytochrome c oxidase subunit II (COII) gene,
            mitochondrial gene encoding mitochondrial protein, partial cds.
ACCESSION   U36846
VERSION     U36846.1  GI:1040987
...
     CDS             <1..>568
                     /gene="COII"
                     /codon_start=1
                     /product="cytochrome c oxidase subunit II"
                     /protein_id = "AAA12345.1"
                     /db_xref="GI:1040988"
```

Note the eventual removal of NID and PID, and the preservation of the ACCESSION linetype. Note also that, if you use NCBI gi identifiers to link to NCBI systems, they will remain available via the VERSION linetype and the type "GI" /db_xref qualifier.

Detailed examples of the new accession number format and the manner in which they will appear in GenBank flatfiles will be provided via upcoming GenBank Release Notes and posts to the bionet.molbio.genbank newsgroup. Subject to synchronization of this change with EMBL and DDBJ, we plan to introduce 'Accession.Version' in 1998.

1.5 Request for Direct Submission of Sequence Data

A successful GenBank requires that the data enter the database as soon as possible after publication, that the annotations be as complete as possible, and that the sequence and annotation data be accurate. All three of these requirements are best met if authors of sequence data submit their data directly to GenBank in a usable form. It is especially important that these submissions be in computer-readable form.

GenBank must rely on direct author submission of data to ensure that it achieves its goals of completeness, accuracy, and timeliness. To assist researchers in entering their own sequence data, GenBank provides a WWW submission tool called BankIt, as well as a stand-alone software package called Sequin. BankIt and Sequin are both easy-to-use programs that enable authors to enter a sequence, annotate it, and submit it to GenBank. Through the international collaboration of DNA sequence databases, GenBank submissions are forwarded daily for inclusion in the EMBL and DDBJ databases.

SEQUIN. Sequin is an interactive, graphically-oriented program based on screen forms and controlled vocabularies that guides you through the process of entering your sequence and providing biological and bibliographic annotation. Intended as an alternative to the older Authorin program, Sequin is designed to simplify the sequence submission process, and to provide increased data handling capabilities to accomodate very long sequences, complex annotations, and robust error checking. E-mail the completed submission file to : gb-sub@ncbi.nlm.nih.gov

Sequin is currently provided as a beta-test version, and runs on

Macintosh, PC/Windows, UNIX and VMS computers. It is available by annonymous ftp from ncbi.nlm.nih.gov , login as anonymous and use your e-mail address as the password. It is located in the sequin directory.

BANKIT. BankIt provides a simple forms approach for submitting your sequence and descriptive information to GenBank. Your submission will be submitted directly to GenBank via the World Wide Web, and immediately forwarded for inclusion in the EMBL and DDBJ databases. BankIt may be used with Netscape clients for Unix, Macs, and PCs, the Mosaic client for Unix, and the MacWeb client for Macs. You can access BankIt from GenBank's home page: http://www.ncbi.nlm.nih.gov/

AUTHORIN. Authorin is no longer the primary means for submitting sequences to GenBank, and is no longer being distributed by NCBI. For submitters who already have this program, however, we do continue to accept Authorin submissions.

For those who are unable to use Sequin or BankIt, GenBank has an ASCII text electronic data submission form. This form is standardized among EMBL, DDBJ, GenBank, PIR, MIPS, and JIPID. The GenBank Data Submission Form (located in the file GBDAT.FRM) can be used to submit your sequence and annotations. Electronic mail submissions should go to: gb-sub@ncbi.nlm.nih.gov. Direct mail on floppy disk should go to:

GenBank Submissions
National Center for Biotechnology Information
Bldg. 38A, Rm 8N-803
8600 Rockville Pike
Bethesda, MD 20894

If you have questions about GenBank submissions or any of the data submission tools, contact NCBI at: info@ncbi.nlm.nih.gov or 301-496-2475.

1.6 Organization of This Document

The second section describes the contents of the CD-ROM files. The third section illustrates the formats of the CD-ROM files. The fourth section describes other versions of the data, the fifth section identifies known problems, and the sixth contains administrative details and ordering information.

2. ORGANIZATION OF CD-ROM FILES

2.1 CD-ROM Format

The GenBank CD-ROM distribution files are available on ISO-9660 compatible CD-ROM. The data are written as ASCII files with variable length records. Each record corresponds to one line in the data bank; a carriage return/line feed pair terminate each line.

The data on the CD-ROMs have both uppercase and lowercase characters.

2.2 Files

The GenBank flat file release consists of forty-four files on the CD-ROM. The list that follows describes each of the files included in the distribution. Their sizes and base pair content are also summarized.

2.2.1 File Descriptions

1. gbrel.txt - Release notes (this document).
2. gbsdr.txt - Short directory of the data bank.
3. gbacc.idx - Index of the entries according to accession number.
4. gbkey.idx - Index of the entries according to keyword phrase.
5. gbaut.idx - Index of the entries according to author.
6. gbjou.idx - Index of the entries according to journal citation.
7. gbgen.idx - Index of the entries according to gene names.
8. gbdat.frm - Forms for submitting sequences or corrections to GenBank.
9. gbpri1.seq - Primate sequence entries, part 1.
10. gbpri2.seq - Primate sequence entries, part 2.
11. gbrod.seq - Rodent sequence entries.
12. gbmam.seq - Other mammalian sequence entries.
13. gbvrt.seq - Other vertebrate sequence entries.
14. gbinv.seq - Invertebrate sequence entries.
15. gbpln.seq - Plant sequence entries (including fungi and algae).
16. gbbct.seq - Bacterial sequence entries.
17. gbrna.seq - Structural RNA sequence entries.
18. gbvrl.seq - Viral sequence entries.
19. gbphg.seq - Phage sequence entries.
20. gbsyn.seq - Synthetic and chimeric sequence entries.
21. gbuna.seq - Unannotated sequence entries.
22. gbest1.seq - EST (expressed sequence tag) sequence entries, part 1.
23. gbest2.seq - EST (expressed sequence tag) sequence entries, part 2.
24. gbest3.seq - EST (expressed sequence tag) sequence entries, part 3.
25. gbest4.seq - EST (expressed sequence tag) sequence entries, part 4.
26. gbest5.seq - EST (expressed sequence tag) sequence entries, part 5.
27. gbest6.seq - EST (expressed sequence tag) sequence entries, part 6.
28. gbest7.seq - EST (expressed sequence tag) sequence entries, part 7.
29. gbest8.seq - EST (expressed sequence tag) sequence entries, part 8.
30. gbest9.seq - EST (expressed sequence tag) sequence entries, part 9.
31. gbest10.seq - EST (expressed sequence tag) sequence entries, part 10.
32. gbest11.seq - EST (expressed sequence tag) sequence entries, part 11.
33. gbest12.seq - EST (expressed sequence tag) sequence entries, part 12.
34. gbest13.seq - EST (expressed sequence tag) sequence entries, part 13.
35. gbest14.seq - EST (expressed sequence tag) sequence entries, part 14.
36. gbest15.seq - EST (expressed sequence tag) sequence entries, part 15.
37. gbest16.seq - EST (expressed sequence tag) sequence entries, part 16.
38. gbest17.seq - EST (expressed sequence tag) sequence entries, part 17.
39. gbest18.seq - EST (expressed sequence tag) sequence entries, part 18.
40. gbest19.seq - EST (expressed sequence tag) sequence entries, part 19.
41. gbpat.seq - Patent sequence entries.
42. gbsts.seq - STS (sequence tagged site) sequence entries.
43. gbgss.seq - GSS (genome survey sequence) sequence entries.
44. gbhtg.seq - HTGS (high throughput genomic sequencing) sequence entries.

2.2.5 File Sizes

The following table indicates the approximate sizes of the individual files in this release. Since minor changes to some of the files may occur after the release notes are written, these sizes should not be used to determine file integrity; they are provided as an aid to planning only. Note also that the sizes of the files in the CdRom distribution are somewhat larger due to the presence of carriage-return and linefeed characters at the end of each line.

File Size	File Name
68386849	gbacc.idx
852219766	gbaut.idx
253774284	gbbct.seq
22904	gbdat.frm

```
177466709       gbest1.seq
230310799       gbest10.seq
242141254       gbest11.seq
201901996       gbest12.seq
183841993       gbest13.seq
222791086       gbest14.seq
220426769       gbest15.seq
186900181       gbest16.seq
214516249       gbest17.seq
211880783       gbest18.seq
 16664425       gbest19.seq
210866853       gbest2.seq
217232974       gbest3.seq
213554250       gbest4.seq
209869937       gbest5.seq
244182651       gbest6.seq
231050931       gbest7.seq
217252533       gbest8.seq
228507630       gbest9.seq
  7153724       gbgen.idx
194651247       gbgss.seq
114530506       gbhtg.seq
220412927       gbinv.seq
 78887509       gbjou.idx
 64209869       gbkey.idx
 40961457       gbmam.seq
106693407       gbpat.seq
  6260115       gbphg.seq
235784470       gbpln.seq
121929981       gbpri1.seq
225094137       gbpri2.seq
    90508       gbrel.txt
  9511785       gbrna.seq
136656785       gbrod.seq
151358535       gbsdr.txt
135695808       gbsts.seq
 12835371       gbsyn.seq
  6462110       gbuna.seq
160498983       gbvrl.seq
 56747841       gbvrt.seq
```

2.2.6 Per-Division Statistics

The following table provides a per-division breakdown of the number of sequence entries and the total number of bases of DNA/RNA in each sequence data file:

Division	Entries	Bases
BCT	40215	99916287
EST1	74012	25280294
EST2	74000	27310278
EST3	74002	27626262
EST4	74000	26781179
EST5	74000	25844882
EST6	74001	27908982
EST7	74000	28338464
EST8	74000	29191562
EST9	74000	28777624
EST10	74000	26953738

EST11	74000	21792490
EST12	74000	27025253
EST13	74000	25063057
EST14	74000	30111226
EST15	74000	27225407
EST16	74000	25801273
EST17	74000	29552886
EST18	74000	31019428
EST19	5656	2100549
GSS	85678	41443623
HTG	1674	86361628
INV	30945	108602308
MAM	13105	12814580
PAT	87767	27593724
PHG	1345	2169977
PLN	46684	97539963
PRI1	40020	37933184
PRI2	39100	110315832
RNA	4789	2469685
ROD	37916	47204565
STS	53287	18376818
SYN	2616	5884669
UNA	2395	1972622
VRL	48503	46327741
VRT	18243	17658473

2.2.7 Selected Per-Organism Statistics

The following table provides the number of entries and bases of DNA/RNA for the twenty most sequenced organisms in Release 104.0 (chloroplast and mitochondrial sequences not included):

Entries	Bases	Species
1038373	551650819	Homo sapiens
281308	133404600	Mus musculus
75960	113659949	Caenorhabditis elegans
56146	46566672	Arabidopsis thaliana
26211	29198918	Drosophila melanogaster
10474	28567944	Saccharomyces cerevisiae
4743	17345841	Escherichia coli
10517	13826989	Rattus norvegicus
1058	9715721	Bacillus subtilis
21006	9198129	Human immunodeficiency virus type 1
22270	8932632	Oryza sativa
15057	7522232	Fugu rubripes
1225	6398746	Schizosaccharomyces pombe
4484	5099500	Gallus gallus
580	4509620	Mycobacterium tuberculosis
10818	4350356	Toxoplasma gondii
11457	4312388	Brugia malayi
4704	4164811	Bos taurus
147	3829215	Synechocystis sp.
2067	3160564	Xenopus laevis

2.2.8 Growth of GenBank

The following table lists the number of bases and the number of sequence records in each release of GenBank, beginning with Release 3 in 1982. Over the period 1982 to the present, the number of bases in GenBank

has doubled approximately every 14 months.

Release	Date	Base Pairs	Entries
3	Dec 82	680338	606
14	Nov 83	2274029	2427
20	May 84	3002088	3665
24	Sep 84	3323270	4135
25	Oct 84	3368765	4175
26	Nov 84	3689752	4393
32	May 85	4211931	4954
36	Sep 85	5204420	5700
40	Feb 86	5925429	6642
42	May 86	6765476	7416
44	Aug 86	8442357	8823
46	Nov 86	9615371	9978
48	Feb 87	10961380	10913
50	May 87	13048473	12534
52	Aug 87	14855145	14020
53	Sep 87	15514776	14584
54	Dec 87	16752872	15465
55	Mar 88	19156002	17047
56	Jun 88	20795279	18226
57	Sep 88	22019698	19044
57.1	Oct 88	23800000	20579
58	Dec 88	24690876	21248
59	Mar 89	26382491	22479
60	Jun 89	31808784	26317
61	Sep 89	34762585	28791
62	Dec 89	37183950	31229
63	Mar 90	40127752	33377
64	Jun 90	42495893	35100
65	Sep 90	49179285	39533
66	Dec 90	51306092	41057
67	Mar 91	55169276	43903
68	Jun 91	65868799	51418
69	Sep 91	71947426	55627
70	Dec 91	77337678	58952
71	Mar 92	83894652	65100
72	Jun 92	92160761	71280
73	Sep 92	101008486	78608
74	Dec 92	120242234	97084
75	Feb 93	126212259	106684
76	Apr 93	129968355	111911
77	Jun 93	138904393	120134
78	Aug 93	147215633	131328
79	Oct 93	157152442	143492
80	Dec 93	163802597	150744
81	Feb 94	173261500	162946
82	Apr 94	180589455	169896
83	Jun 94	191393939	182753
84	Aug 94	201815802	196703
85	Oct 94	217102462	215273
86	Dec 94	230485928	237775
87	Feb 95	248499214	269478
88	Apr 95	286094556	352414
89	Jun 95	318624568	425211
90	Aug 95	353713490	492483
91	Oct 95	384939485	555694

92	Dec 95	425860958	620765
93	Feb 96	463758833	685693
94	Apr 96	499127741	744295
95	Jun 96	551750920	835487
96	Aug 96	602072354	920588
97	Oct 96	651972984	1021211
98	Dec 96	730552938	1114581
99	Feb 97	786898138	1192505
100	Apr 97	842864309	1274747
101	Jun 97	966993087	1491069
102	Aug 97	1053474516	1610848
103	Oct 97	1160300687	1765847
104	Dec 97	1258290513	1891953

3. FILE FORMATS

The flat file examples included in this section, while not always from the current release, are usually quite recent. Any differences compared to the actual data files are the result of updates to the entries involved.

3.1 File Header Information

Each of the forty-four files on the distribution CD-ROM begins with the same header, except for the first line, which contains the file name, and the sixth line, which contains the title of the file. The first line of the file contains the file name in character positions 1 to 9 and the full data bank name (Genetic Sequence Data Bank) starting in column 20. The brief names of the files in this release are listed in section 2.2.

The second line contains the date of the current release in the form `day month year', beginning in position 26. The fourth line contains the current GenBank release number. The release number appears in positions 41 to 45 and consists of two numbers separated by a decimal point. The number to the left of the decimal is the major release number. The digit to the right of the decimal indicates the version of the major release; it is zero for the first version. The sixth line contains a title for the file. The eighth line lists the number of entries (loci), number of bases (or base pairs), and number of reports of sequences (equal to number of entries in this case). These numbers are right-justified at fixed positions. The number of entries appears in positions 1 to 7, the number of bases in positions 15 to 23, and the number of reports in positions 37 to 40. (There are more reports of sequences than entries since reported sequences that overlap or duplicate each other are combined into single entries.) The third, fifth, seventh, and ninth lines are blank.

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
GBACC.IDX          Genetic Sequence Data Bank
                         15 December 1993

                  GenBank Flat File Release 80.0

                      Accession Number Index

  150744 loci, 163802597 bases, from 150744 reported sequences
--------+---------+---------+---------+---------+---------+---------+---------
1       10        20        30        40        50        60        70        79
```

Example 1. Sample File Header

3.2 Directory Files

3.2.1 Short Directory File

The short directory file contains brief descriptions of all of the sequence entries contained in this release. These descriptions are in fifteen groups, one group for each of the fifteen sequence entry data files. The first record at the beginning of a group of entries contains the name of the group in uppercase characters, beginning in position 21. The organism groups are PRIMATE, RODENT, OTHER MAMMAL, OTHER VERTEBRATE, INVERTEBRATE, PLANT, BACTERIAL, STRUCTURAL RNA, VIRAL, PHAGE, SYNTHETIC, UNANNOTATED, EXPRESSED SEQUENCE TAG, PATENT, or SEQUENCE TAGGED SITE. The second record is blank.

Each record in the short directory contains the sequence entry name (LOCUS) in the first 12 positions, followed by a brief definition of the sequence beginning in column 13. The definition is truncated (at the end of a word) to leave room at the right margin for at least one space, the sequence length, and the letters `bp'. The length of the sequence is printed right-justified to column 77, followed by the letters `bp' in columns 78 and 79. The next-to-last record for a group has `ZZZZZZZZZZ' in its first ten positions (where the entry name would normally appear). The last record is a blank line. An example of the short directory file format, showing the descriptions of the last entries in the Other Vertebrate sequence data file and the first entries of the Invertebrate sequence data file, is reproduced below:

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
ZEFWNT1G3   B.rerio wnt-1 gene (exon 3) for wnt-1 protein.                 266bp
ZEFWNT1G4   B.rerio wnt-1 gene (exon 4) for wnt-1 protein.                 647bp
ZEFZF54     Zebrafish homeotic gene ZF-54.                                246bp
ZEFZFEN     Zebrafish engrailed-like homeobox sequence.                   327bp
ZZZZZZZZZZ

                    INVERTEBRATE

AAHAV33A    Acanthocheilonema viteae pepsin-inhibitor-like-protein       1048bp
ACAAC01     Acanthamoeba castelani gene encoding actin I.                1571bp
ACAACTPH    Acanthamoeba castellanii actophorin mRNA, complete cds.      671bp
ACAMHCA     A.castellanii non-muscle myosin heavy chain gene, partial    5894bp
--------+---------+---------+---------+---------+---------+---------+---------
1       10        20        30        40        50        60        70        79
```

Example 2. Short Directory File

3.3 Index Files

There are five files containing indices to the entries in this release:

Accession number index file
Keyword phrase index file
Author name index file
Journal citation index file
Gene name index file

The index keys (accession numbers, keywords, authors, journals, and

gene symbols.) of an index are sorted alphabetically. (The index keys for the keyword phrases and author names appear in uppercase characters even though they appear in mixed case in the sequence entries.) Under each index key, the names of the sequence entries containing that index key are listed alphabetically. Each sequence name is also followed by its data file division and primary accession number. The following codes are used to designate the data file divisions:

1. PRI - primate sequences
2. ROD - rodent sequences
3. MAM - other mammalian sequences
4. VRT - other vertebrate sequences
5. INV - invertebrate sequences
6. PLN - plant, fungal, and algal sequences
7. BCT - bacterial sequences
8. RNA - structural RNA sequences
9. VRL - viral sequences
10. PHG - bacteriophage sequences
11. SYN - synthetic sequences
12. UNA - unannotated sequences
13. EST - EST sequences (expressed sequence tags)
14. PAT - patent sequences
15. STS - STS sequences (sequence tagged sites)
16. GSS - GSS sequences (genome survey sequences)
17. HTG - HTGS sequences (high throughput genomic sequences)

The index key begins in column 1 of a record. An 11-character field for the sequence entry name starts in position 14 of a record, followed by a 3-character field for the data file division, starting at position 25 and ending at position 27, and a 6-character field for the primary accession number, starting at position 29 and ending at position 34. All entries in the fields are left-justified.

Beginning at positions 36 and 58, the three fields repeat, so three sets of sequence information can appear in one record. If there are more than three entry names, the next records are used; the index key is not repeated. For the accession number files, the entry names begin in the same record as the index key, since the key is always less than 12 characters. In the other index files, the entry names begin on the record following the index key record.

NOTE: The column positions stated above will be shifted to the right if primary accessions in the 8-character format are present.

3.3.1 Accession Number Index File

Accession numbers are unique six character or eight-character alphanumeric identifiers of GenBank database entries. The six-character accession number format consists of a single uppercase letter, followed by 5 digits. The eight-character accession number format consists of two uppercase letters, followed by 6 digits. Accessions provide an unchanging identifier for the data with which they are associated, and we encourage you to cite accession numbers whenever you refer to data from GenBank.

GenBank entries can have both 'primary' and 'secondary' accessions associated with them (see Section 3.5.6).

The following excerpt from the accession number index file illustrates the format of the index:

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
J00316       HUMTBB11P  PRI J00316
J00317       HUMTBB46P  PRI J00317
J00318       HUMUG1     PRI J00318
J00319       HUMUG1PA   PRI J00319
J00320       HUMVIPMR1  PRI L00154 HUMVIPMR2  PRI L00155 HUMVIPMR3  PRI L00156
             HUMVIPMR4  PRI L00157 HUMVIPMR5  PRI L00158
J00321       BABA1AT    PRI J00321
J00322       CHPRSA     PRI J00322
J00323       AGMRSASPC  PRI J00323
J00324       BABATIII   PRI J00324
--------+---------+---------+---------+---------+---------+---------+---------
1       10        20        30        40        50        60        70        79
```

Example 4. Accession Number Index File

If the same accession number is found in more than one entry (a result of the infrequent occasions when a single entry is split into two or more separate entries), then the additional entries and groups in which the number appears are also given. In the example above, J00320 is a secondary accession, appearing on 5 other database entries.

3.3.2 Keyword Phrase Index File

Keyword phrases consist of names for gene products and other characteristics of sequence entries. There are approximately 18,000 keyword phrases. An excerpt from the keyword phrase index file is shown below:

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
DNA HELICASE
             ECOHELIV   BCT J04726 ECOUVRD    BCT X00738 FPLTRAX    BCT M38047
             HS2ULL     VRL D10470 HSECOMGEN  VRL M86664 PT4DDA     PHG M93048
             SYNPMMB190 SYN M37846 YSPRHP3    PLN X64583
DNA HELICASE I
             ECOPTRAI5  BCT X57430
DNA HELICASE II
             ECOUVRD2   BCT D00069 HEAMUTB1A  BCT M99049
DNA INVERSION SYSTEM
             ECOP15BG   BCT X62121
DNA INVERTASE
             ECOPIN     BCT K00676 ECOPIN1    BCT X01805 PMUGINMOM  PHG V01463
             STABINR3   BCT X16298 STAINVSA   BCT M36694
DNA J HEATSHOCK PROTEIN
             MSGDNAJHSP BCT M95576
DNA LIGASE
             ECOLIG     BCT M24278 ECOLIGA    BCT M30255 PT4G30     PHG X00039
             PT6LIG55   PHG M38465 TTHDNALGS  BCT M74792 TTHDNALIG  BCT M36417
             VACCDNLIG  VRL X16512 VACRHF     VRL D11079 YSCCDC9    PLN X03246
             YSPCDC17   PLN X05107 ZMOLIG     BCT Z11910
--------+---------+---------+---------+---------+---------+---------+---------
1       10        20        30        40        50        60        70        79
```

Example 5. Keyword Phrase Index File

3.3.3 Author Name Index File

The author name index file lists all of the author names that appear in the citations. An excerpt from the author name index file is shown below:

```
1        10        20        30        40        50        60        70        79
---------+---------+---------+---------+---------+---------+---------+---------
JACKSON,D.I.
              RATLCAG1    ROD M18349 RATLCAG2    ROD M18348 RATLCAG3    ROD M18347
              RATLCAI     ROD M25820 RATLCAII    ROD M25821 RATLCAIII   ROD M25822
              RATLCAIV    ROD M25823 RATLCAR     ROD Y00065
JACKSON,F.R.
              DRO16883C   INV X62939 DRO1688ED   INV X62938 DRO1688EP   INV X62937
              DROPER      INV M11969 DROPES      INV X03636 MUSPER      ROD M12039
              MUSURFPER   ROD X02966
JACKSON,I.J.
              MUSHOMA     ROD X03033 MUSNEORP8R  ROD X54812 MUSP7H2     ROD X54811
              MUSRPT      ROD M69041 MUSSOFI     ROD X63350 MUSTRP15    ROD X59513
              MUSTYRP2    ROD X63349
JACKSON,I.M.
              RATTRH      ROD M12138
JACKSON,J.
              DROFPS85D   INV X52844 MUSIGKAC3   ROD K00885 MUSIL4RA    ROD M27959
              MUSIL4RB    ROD M27960 RABGLOBCON  MAM L05833 RABGLOBHSB  MAM L05835
---------+---------+---------+---------+---------+---------+---------+---------
1        10        20        30        40        50        60        70        79
```

Example 6. Author Name Index File

3.3.4 Journal Citation Index File

The journal citation index file lists all of the citations that appear in the references. All citations are truncated to 80 characters. An excerpt from the citation index file is shown below:

```
1        10        20        30        40        50        60        70        79
---------+---------+---------+---------+---------+---------+---------+---------
(IN) THE IMMUNE SYSTEM:  132-138, S. KARGER, NEW YORK (1981)
              HUMIGHVX    PRI M35415
(IN) THE LENS: TRANSPARANCY AND CATARACT:  171-179, EURAGE, RIJSWIJK (1986)
              RANCRYG2A   VRT K02264 RANCRYG4A   VRT K02266 RANCRYG5A   VRT M22529
              RANCRYG6A   VRT M22530 RANCRYR     VRT X00659
(IN) THIOREDOXIN AND GLUTAREDOXIN SYSTEMS: STRUCTURE AND FUNCTION: 11-19, UNKNOW
              ECOTRXA1    BCT M54881
(IN) UCLA SYMP. MOL. CELL. BIOL. NEW SER., VOL. 77:  339-352, ALAN R. LISS, INC.
              BOVTRNB2A   MAM M36431 HUMTRNB     PRI M36429 HUMTRNB1    PRI M36430
(IN) UCLA SYMPOSIA:  575-584, ALAN R. LISS, INC., NEW YORK (1987)
              PFAHGPRT    INV M54896
(IN) VIRUS RESEARCH. PROCEEDINGS OF 1973 ICN-UCLA SYMPOSIUM:  533-544, ACADEMIC
              LAMCG       PHG J02459
ACTA BIOCHIM. BIOPHYS. SIN. 23, 246-253 (1992)
              HUMPLASINS  PRI M98056
---------+---------+---------+---------+---------+---------+---------+---------
1        10        20        30        40        50        60        70        79
```

Example 7. Journal Citation Index File

3.3.5 Gene Name Index

The /gene qualifiers of many GenBank entries contain values other than official gene symbols, such as the product or the standard name of the gene. Hence, NCBI has chosen to build an index (gbgen.idx) more like a keyword index for this field, using both the GenBank /gene qualifier and the 'Gene.locus' fields from the NCBI internal database as keys. An excerpt from the gene name index file is shown below:

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
SUPPRESSOR OF SABLE
            DROSUSG    INV M57889
SUPPRESSOR TWO OF ZESTE
            DROS2ZSTG  INV X56798
SUPRESSOR TWO OF ZESTE
            DROS2ZSTM  INV X56799
SUR
            CHKSRVCNTK VRT M57290
SURC
            ARFSURCG   BCT X63435
--------+---------+---------+---------+---------+---------+---------+---------
1       10        20        30        40        50        60        70        79
```

Example 8. Gene Name Index File

3.4 GenBank Data Submission Form and Error/Suggestion Report Form

The recommended methods for submitting sequence data to GenBank are via BankIt and the Sequin program. Please see Section 1.5 of this document for further details.

If it is not possible to use BankIt or Sequin, there is a data submission form in this distribution (GBDAT.FRM) which can be filled out with a text editor and returned to the database, preferably by e-mail.

Direct submission e-mail address: gb-sub@ncbi.nlm.nih.gov

The second form in the GBDAT.FRM is the GenBank Error/Suggestion Report Form. It is separated from the Data Submission Form by a form-feed character (<CTRL>L, ASCII octal value 014, ASCII decimal value 12). We encourage all users to report any errors to the data bank staff using this form. Like the GenBank Data Submission Form, it may be printed and filled in by hand and sent by mail to the address given at the beginning of the form. It may also be filled out using a text editor and sent to GenBank by electronic mail at: update@ncbi.nlm.nih.gov

3.5 Sequence Entry Files

The distribution CD-ROM contains fifteen sequence entry data files, one for each division of GenBank.

3.5.1 File Organization

Each of these files has the same format and consists of two parts: header information (described in section 3.1) and sequence entries for that division (described in the following sections).

3.5.2 Entry Organization

In the second portion of a sequence entry file (containing the sequence entries for that division), each record (line) consists of two parts. The first part is found in positions 1 to 10 and may contain:

1. A keyword, beginning in column 1 of the record (e.g., REFERENCE is a keyword).

2. A subkeyword beginning in column 3, with columns 1 and 2 blank (e.g., AUTHORS is a subkeyword of REFERENCE).

3. Blank characters, indicating that this record is a continuation of the information under the keyword or subkeyword above it.

4. A code, beginning in column 5, indicating the nature of an entry (feature key) in the FEATURES table; these codes are described in Section 3.5.12.1 below.

5. A number, ending in column 9 of the record. This number occurs in the portion of the entry describing the actual nucleotide sequence and designates the numbering of sequence positions.

6. Two slashes (//) in positions 1 and 2, marking the end of an entry.

The second part of each sequence entry record contains the information appropriate to its keyword, in positions 13 to 80 for keywords and positions 11 to 80 for the sequence.

The following is a brief description of each entry field. Detailed information about each field may be found in Sections 3.5.4 to 3.5.14.

LOCUS - A short mnemonic name for the entry, chosen to suggest the sequence's definition. Mandatory keyword/exactly one record.

DEFINITION - A concise description of the sequence. Mandatory keyword/one or more records.

ACCESSION - The primary accession number is a unique, unchanging code assigned to each entry. (Please use this code when citing information from GenBank.) Mandatory keyword/one or more records.

NID - The unique nucleic acid identifier that has been assigned to the current version of the sequence data that are associated with the GenBank entry identified by a given primary accession number.

KEYWORDS - Short phrases describing gene products and other information about an entry. Mandatory keyword in all annotated entries/one or more records.

SEGMENT - Information on the order in which this entry appears in a series of discontinuous sequences from the same molecule. Optional keyword (only in segmented entries)/exactly one record.

SOURCE - Common name of the organism or the name most frequently used in the literature. Mandatory keyword in all annotated entries/one or more records/includes one subkeyword.

ORGANISM - Formal scientific name of the organism (first line) and taxonomic classification levels (second and subsequent lines).

Mandatory subkeyword in all annotated entries/two or more records.

REFERENCE - Citations for all articles containing data reported in this entry. Includes four subkeywords and may repeat. Mandatory keyword/one or more records.

AUTHORS - Lists the authors of the citation. Mandatory subkeyword/one or more records.

TITLE - Full title of citation. Optional subkeyword (present in all but unpublished citations)/one or more records.

JOURNAL - Lists the journal name, volume, year, and page numbers of the citation. Mandatory subkeyword/one or more records.

MEDLINE - Provides the Medline unique identifier for a citation. Optional subkeyword/one record.

REMARK - Specifies the relevance of a citation to an entry. Optional subkeyword/one or more records.

COMMENT - Cross-references to other sequence entries, comparisons to other collections, notes of changes in LOCUS names, and other remarks. Optional keyword/one or more records/may include blank records.

FEATURES - Table containing information on portions of the sequence that code for proteins and RNA molecules and information on experimentally determined sites of biological significance. Optional keyword/one or more records.

BASE COUNT - Summary of the number of occurrences of each base code in the sequence. Mandatory keyword/exactly one record.

ORIGIN - Specification of how the first base of the reported sequence is operationally located within the genome. Where possible, this includes its location within a larger genetic map. Mandatory keyword/exactly one record.

- The ORIGIN line is followed by sequence data (multiple records).

// - Entry termination symbol. Mandatory at the end of an entry/exactly one record.

3.5.3 Sample Sequence Data File

An example of a complete sequence entry file follows. (This example has only two entries.) Note that in this example, as throughout the data bank, numbers in square brackets indicate items in the REFERENCE list. For example, in ACARR58S, [1] refers to the paper by Mackay, et al.

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
GBSMP.SEQ          Genetic Sequence Data Bank
                         15 December 1992

                  GenBank Flat File Release 74.0

                     Structural RNA Sequences
```

```
     2 loci,       236 bases, from     2 reported sequences

LOCUS       AAURRA        118 bp ss-rRNA            RNA       16-JUN-1986
DEFINITION  A.auricula-judae (mushroom) 5S ribosomal RNA.
ACCESSION   K03160
KEYWORDS    5S ribosomal RNA; ribosomal RNA.
SOURCE      A.auricula-judae (mushroom) ribosomal RNA.
  ORGANISM  Auricularia auricula-judae
            Eukaryota; Fungi; Eumycota; Basidiomycotina; Phragmobasidiomycetes;
            Heterobasidiomycetidae; Auriculariales; Auriculariaceae.
REFERENCE   1  (bases 1 to 118)
  AUTHORS   Huysmans,E., Dams,E., Vandenberghe,A. and De Wachter,R.
  TITLE     The nucleotide sequences of the 5S rRNAs of four mushrooms and
            their use in studying the phylogenetic position of basidiomycetes
            among the eukaryotes
  JOURNAL   Nucleic Acids Res. 11, 2871-2880 (1983)
FEATURES             Location/Qualifiers
     rRNA            1..118
                     /note="5S ribosomal RNA"
BASE COUNT       27 a     34 c     34 g     23 t
ORIGIN      5' end of mature rRNA.
        1 atccacggcc ataggactct gaaagcactg catcccgtcc gatctgcaaa gttaaccaga
       61 gtaccgccca gttagtacca cggtgggggа ccacgcggga atcctgggtg ctgtggtt
//
LOCUS       ABCRRAA       118 bp ss-rRNA            RNA       15-SEP-1990
DEFINITION  Acetobacter sp. (strain MB 58) 5S ribosomal RNA, complete sequence.
ACCESSION   M34766
KEYWORDS    5S ribosomal RNA.
SOURCE      Acetobacter sp. (strain MB 58) rRNA.
  ORGANISM  Acetobacter sp.
            Prokaryotae; Gracilicutes; Scotobacteria; Aerobic rods and cocci;
            Azotobacteraceae.
REFERENCE   1  (bases 1 to 118)
  AUTHORS   Bulygina,E.S., Galchenko,V.F., Govorukhina,N.I., Netrusov,A.I.,
            Nikitin,D.I., Trotsenko,Y.A. and Chumakov,K.M.
  TITLE     Taxonomic studies of methylotrophic bacteria by 5S ribosomal RNA
            sequencing
  JOURNAL   J. Gen. Microbiol. 136, 441-446 (1990)
FEATURES             Location/Qualifiers
     rRNA            1..118
                     /note="5S ribosomal RNA"
BASE COUNT       27 a     40 c     32 g     17 t      2 others
ORIGIN
        1 gatctggtgg ccatggcggg agcaaatcag ccgatcccat cccgaactcg gccgtcaaat
       61 gccccagcgc ccatgatact ctgcctcaag gcacggaaaa gtcggtcgcc gccagayy
//
---------+---------+---------+---------+---------+---------+---------+---------
1        10        20        30        40        50        60        70        79
```

Example 9. Sample Sequence Data File

3.5.4 LOCUS Format

The pieces of information contained in the LOCUS record are always found in fixed positions. The locus name (or entry name), which is always ten characters or less, begins in position 13. The locus name is designed to help group entries with similar sequences: the first three characters usually designate the organism; the fourth and fifth characters can be used to show other group designations, such as gene

product; for segmented entries the last character is one of a series of sequential integers.

The number of bases or base pairs in the sequence ends in position 29. The letters `bp' are in positions 31 to 32. Positions 34 to 36 give the number of strands of the sequence. Positions 37 to 40 give the topology of molecule sequenced. If the sequence is of a special type, a notation (such as `circular') is included in positions 43 to 52.

GenBank sequence entries are divided among fifteen taxonomic divisions. Each entry's division is identified by a three-letter code in positions 53 to 55. See Section 3.3 for the division codes.

Positions 63 to 73 of the record contain the date the entry was entered or underwent any substantial revisions, such as the addition of newly published data, in the form dd-MMM-yyyy.

The detailed format for the LOCUS record is as follows:

Positions	Contents
1-12	LOCUS
13-22	Locus name
23-29	Length of sequence, right-justified
31-32	bp
34-36	Blank, ss- (single-stranded), ds- (double-stranded), or ms- (mixed-stranded)
37-40	Blank, DNA, RNA, tRNA (transfer RNA), rRNA (ribosomal RNA), mRNA (messenger RNA), or uRNA (small nuclear RNA)
43-52	Blank (implies linear) or circular
53-55	The division code (see Section 3.3)
63-73	Date, in the form dd-MMM-yyyy (e.g., 15-MAR-1991)

3.5.5 DEFINITION Format

The DEFINITION record gives a brief description of the sequence, proceeding from general to specific. It starts with the common name of the source organism, then gives the criteria by which this sequence is distinguished from the remainder of the source genome, such as the gene name and what it codes for, or the protein name and mRNA, or some description of the sequence's function (if the sequence is non-coding). If the sequence has a coding region, the description may be followed by a completeness qualifier, such as cds (complete coding sequence). There is no limit on the number of lines that may be part of the DEFINITION. The last line must end with a period.

3.5.5.1 DEFINITION Format for NLM Entries

The DEFINITION line for entries derived from journal-scanning at the NLM is an automatically generated descriptive summary that accompanies each DNA and protein sequence. It contains information derived from fields in a database that summarize the most important attributes of the sequence. The DEFINITION lines are designed to supplement the accession number and the sequence itself as a means of uniquely and completely specifying DNA and protein sequences. The following are examples of NLM DEFINITION lines:

NADP-specific isocitrate dehydrogenase [swine, mRNA, 1 gene, 1585 nt]

94 kda fiber cell beaded-filament structural protein [rats, lens, mRNA

Partial, 1 gene, 1873 nt]

inhibin alpha {promoter and exons} [mice, Genomic, 1 gene, 1102 nt, segment 1 of 2]

cefEF, cefG=acetyl coenzyme A:deacetylcephalosporin C o-acetyltransferase [Acremonium chrysogenum, Genomic, 2 genes, 2639 nt]

myogenic factor 3, qmf3=helix-loop-helix protein [Japanese quails, embryo, Peptide Partial, 246 aa]

The first part of the definition line contains information describing the genes and proteins represented by the molecular sequences. This can be gene locus names, protein names and descriptions that replace or augment actual names. Gene and gene product are linked by "=". Any special identifying terms are presented within brackets, such as: {promoter}, {N-terminal}, {EC 2.13.2.4}, {alternatively spliced}, or {3' region}.

The second part of the definition line is delimited by square brackets, '[]', and provides details about the molecule type and length. The biological source, i.e., genus and species or common name as cited by the author. Developmental stage, tissue type and strain are included if available. The molecule types include: Genomic, mRNA, Peptide. and Other Genomic Material. Genomic molecules are assumed to be partial sequence unless "Complete" is specified, whereas mRNA and peptide molecules are assumed to be complete unless "Partial" is noted.

3.5.6 ACCESSION Format

This field contains a series of six-character and/or eight-character identifiers called 'accession numbers'. The six-character accession number format consists of a single uppercase letter, followed by 5 digits. The eight-character accession number format consists of two uppercase letters, followed by 6 digits. The 'primary', or first, of the accession numbers occupies positions 13 to 18 (6-character format) or positions 13 to 20 (8-character format). Subsequent 'secondary' accession numbers (if present) are separated from the primary, and from each other, by a single space. In some cases, multiple lines of secondary accession numbers might be present, starting at position 13.

The primary accession number of a GenBank entry provides a stable identifier for the biological object that the entry represents. Accessions do not change when the underlying sequence data or associated features change.

Secondary accession numbers arise for a number of reasons. For example, a single accession number may initially be assigned to a sequence from a publication. If it is later discovered that the sequence must be entered into the database as multiple entries, each entry would receive a new primary accession number, and the original accession number would appear as a secondary accession number on each of the new entries.

3.5.7 NID Format

This field contains the unique nucleic acid sequence identifier that is assigned by NCBI to the sequence data in a GenBank entry. Nucleic acid identifiers consist of the letter 'g', followed by one or more digits. This sequence identifier occupies positions 13 and higher.

While accession numbers allow one to retrieve the same biological entry

in the database, regardless of changes to that record, the nucleic acid identifier changes every time that the underlying sequence changes. Reasons for sequence changes include: the removal of vector contamination, re-sequencing of stretches of ambiguous sequence, and the correction of sequencing errors.

At the NCBI, these nucleic acid sequence identifiers are called "gi" identifiers. In maintaining GenBank, NCBI generates a new gi if a sequence has changed, and then creates pointers between the old and new gis. Retrieval of the particular version of a sequence associated with a gi will always be possible.

3.5.8 KEYWORDS Format

The KEYWORDS field does not appear in unannotated entries, but is required in all annotated entries. Keywords are separated by semicolons; a "keyword" may be a single word or a phrase consisting of several words. Each line in the keywords field ends in a semicolon; the last line ends with a period. If no keywords are included in the entry, the KEYWORDS record contains only a period.

3.5.9 SEGMENT Format

The SEGMENT keyword is used when two (or more) entries of known relative orientation are separated by a short (<10 kb) stretch of DNA. It is limited to one line of the form `n of m', where `n' is the segment number of the current entry and `m' is the total number of segments.

3.5.10 SOURCE Format

The SOURCE field consists of two parts. The first part is found after the SOURCE keyword and contains free-format information including an abbreviated form of the organism name followed by a molecule type; multiple lines are allowed, but the last line must end with a period. The second part consists of information found after the ORGANISM subkeyword. The formal scientific name for the source organism (genus and species, where appropriate) is found on the same line as ORGANISM. The records following the ORGANISM line list the taxonomic classification levels, separated by semicolons and ending with a period.

3.5.11 REFERENCE Format

The REFERENCE field consists of five parts: the keyword REFERENCE, and the subkeywords AUTHORS, TITLE (optional), JOURNAL, MEDLINE (optional), and REMARK (optional).

The REFERENCE line contains the number of the particular reference and (in parentheses) the range of bases in the sequence entry reported in this citation. Additional prose notes may also be found within the parentheses. The numbering of the references does not reflect publication dates or priorities.

The AUTHORS line lists the authors in the order in which they appear in the cited article. Last names are separated from initials by a comma (no space); there is no comma before the final `and'. The list of authors ends with a period. The TITLE line is an optional field, although it appears in the majority of entries. It does not appear in unpublished sequence data entries that have been deposited directly into the GenBank data bank, the EMBL Nucleotide Sequence Data Library,

or the DNA Data Bank of Japan. The TITLE field does not end with a period.

The JOURNAL line gives the appropriate literature citation for the sequence in the entry. The word `Unpublished' will appear after the JOURNAL subkeyword if the data did not appear in the scientific literature, but was directly deposited into the data bank. For published sequences the JOURNAL line gives the Thesis, Journal, or Book citation, including the year of publication, the specific citation, or In press.

The MEDLINE line provides the National Library of Medicine's Medline unique identifier for a citation (if known). Medline UIs are 8 digit numbers.

The REMARK line is a textual comment that specifies the relevance of the citation to the entry.

3.5.12 FEATURES Format

GenBank releases use a feature table format designed jointly by GenBank, the EMBL Nucleotide Sequence Data Library, and the DNA Data Bank of Japan. This format is in use by all three databases. The most complete and accurate Feature Table documentation can be found on the Web at http://www.ncbi.nlm.nih.gov/collab/FT/index.html .

The Feature Table specification is also available as a printed document: `The DDBJ/EMBL/GenBank Feature Table: Definition'. Contact GenBank at the address shown on the first page of these Release Notes if you would like a copy.

The feature table contains information about genes and gene products, as well as regions of biological significance reported in the sequence. The feature table contains information on regions of the sequence that code for proteins and RNA molecules. It also enumerates differences between different reports of the same sequence, and provides cross-references to other data collections, as described in more detail below.

The first line of the feature table is a header that includes the keyword `FEATURES' and the column header `Location/Qualifier.' Each feature consists of a descriptor line containing a feature key and a location (see sections below for details). If the location does not fit on this line, a continuation line may follow. If further information about the feature is required, one or more lines containing feature qualifiers may follow the descriptor line.

The feature key begins in column 6 and may be no more than 15 characters in length. The location begins in column 22. Feature qualifiers begin on subsequent lines at column 22. Location, qualifier, and continuation lines may extend from column 22 to 80.

Feature tables are required, due to the mandatory presence of the source feature. The sections below provide a brief introduction to the feature table format.

3.5.12.1 Feature Key Names

The first column of the feature descriptor line contains the feature key. It starts at column 6 and can continue to column 20. The list of

valid feature keys is shown below.

allele	Related strain contains alternative gene form
attenuator	Sequence related to transcription termination
C_region	Span of the C immunological feature
CAAT_signal	`CAAT box' in eukaryotic promoters
CDS	Sequence coding for amino acids in protein (includes stop codon)
cellular	Region of cellular DNA
conflict	Independent determinations differ
D-loop	Displacement loop
D_region	Span of the D immunological feature
enhancer	Cis-acting enhancer of promoter function
exon	Region that codes for part of spliced mRNA
GC_signal	`GC box' in eukaryotic promoters
iDNA	Intervening DNA eliminated by recombination
insertion_seq	Insertion sequence (IS), a small transposon
intron	Transcribed region excised by mRNA splicing
J_region	Span of the J immunological feature
LTR	Long terminal repeat
mat_peptide	Mature peptide coding region (does not include stop codon)
misc_binding	Miscellaneous binding site
misc_difference	Miscellaneous difference feature
misc_feature	Region of biological significance that cannot be described by any other feature
misc_recomb	Miscellaneous recombination feature
misc_RNA	Miscellaneous transcript feature not defined by other RNA keys
misc_signal	Miscellaneous signal
misc_structure	Miscellaneous DNA or RNA structure
modified_base	The indicated base is a modified nucleotide
mRNA	Messenger RNA
mutation	A mutation alters the sequence here
N_region	Span of the N immunological feature
old_sequence	Presented sequence revises a previous version
polyA_signal	Signal for cleavage & polyadenylation
polyA_site	Site at which polyadenine is added to mRNA
precursor_RNA	Any RNA species that is not yet the mature RNA product
prim_transcript	Primary (unprocessed) transcript
primer	Primer binding region used with PCR
primer_bind	Non-covalent primer binding site
promoter	A region involved in transcription initiation
protein_bind	Non-covalent protein binding site on DNA or RNA
provirus	Proviral sequence
RBS	Ribosome binding site
rep_origin	Replication origin for duplex DNA
repeat_region	Sequence containing repeated subsequences
repeat_unit	One repeated unit of a repeat_region
rRNA	Ribosomal RNA
S_region	Span of the S immunological feature
satellite	Satellite repeated sequence
scRNA	Small cytoplasmic RNA
sig_peptide	Signal peptide coding region
snRNA	Small nuclear RNA
stem_loop	Hair-pin loop structure in DNA or RNA
STS	Sequence Tagged Site; operationally unique sequence that identifies the combination of primer spans used in a PCR assay
TATA_signal	`TATA box' in eukaryotic promoters
terminator	Sequence causing transcription termination
transit_peptide	Transit peptide coding region
transposon	Transposable element (TN)

tRNA	Transfer RNA
unsure	Authors are unsure about the sequence in this region
V_region	Span of the V immunological feature
variation	A related population contains stable mutation
virion	Virion (encapsidated) viral sequence
- (hyphen)	Placeholder
-10_signal	`Pribnow box' in prokaryotic promoters
-35_signal	`-35 box' in prokaryotic promoters
3'clip	3'-most region of a precursor transcript removed in processing
3'UTR	3' untranslated region (trailer)
5'clip	5'-most region of a precursor transcript removed in processing
5'UTR	5' untranslated region (leader)

3.5.12.2 Feature Location

The second column of the feature descriptor line designates the location of the feature in the sequence. The location descriptor begins at position 22. Several conventions are used to indicate sequence location.

Base numbers in location descriptors refer to numbering in the entry, which is not necessarily the same as the numbering scheme used in the published report. The first base in the presented sequence is numbered base 1. Sequences are presented in the 5 to 3 direction.

Location descriptors can be one of the following:

1. A single base;

2. A contiguous span of bases;

3. A site between two bases;

4. A single base chosen from a range of bases;

5. A single base chosen from among two or more specified bases;

6. A joining of sequence spans;

7. A reference to an entry other than the one to which the feature belongs (i.e., a remote entry), followed by a location descriptor referring to the remote sequence;

8. A literal sequence (a string of bases enclosed in quotation marks).

A site between two residues, such as an endonuclease cleavage site, is indicated by listing the two bases separated by a carat (e.g., 23^24).

A single residue chosen from a range of residues is indicated by the number of the first and last bases in the range separated by a single period (e.g., 23.79). The symbols < and > indicate that the end point of the range is beyond the specified base number.

A contiguous span of bases is indicated by the number of the first and last bases in the range separated by two periods (e.g., 23..79). The symbols < and > indicate that the end point of the range is beyond the specified base number. Starting and ending positions can be indicated by base number or by one of the operators described below.

Operators are prefixes that specify what must be done to the indicated sequence to locate the feature. The following are the operators available, along with their most common format and a description.

complement (location): The feature is complementary to the location indicated. Complementary strands are read 5 to 3.

join (location, location, .. location): The indicated elements should be placed end to end to form one contiguous sequence.

order (location, location, .. location): The elements are found in the specified order in the 5 to 3 direction, but nothing is implied about the rationality of joining them.

group (location, location, .. location): The elements are related and should be grouped together, but no order is implied.

one-of (location, location, .. location): The element can be any one, but only one, of the items listed.

3.5.12.3 Feature Qualifiers

Qualifiers provide additional information about features. They take the form of a slash (/) followed by a qualifier name and, if applicable, an equal sign (=) and a qualifier value. Feature qualifiers begin at column 22.

Qualifiers convey many types of information. Their values can, therefore, take several forms:

1. Free text;
2. Controlled vocabulary or enumerated values;
3. Citations or reference numbers;
4. Sequences;
5. Feature labels.

Text qualifier values must be enclosed in double quotation marks. The text can consist of any printable characters (ASCII values 32-126 decimal). If the text string includes double quotation marks, each set must be `escaped' by placing a double quotation mark in front of it (e.g., /note="This is an example of ""escaped"" quotation marks").

Some qualifiers require values selected from a limited set of choices. For example, the `/direction' qualifier has only three values `left,' `right,' or `both.' These are called controlled vocabulary qualifier values. Controlled qualifier values are not case sensitive; they can be entered in any combination of upper- and lowercase without changing their meaning.

Citation or published reference numbers for the entry should be enclosed in square brackets ([]) to distinguish them from other numbers.

A literal sequence of bases (e.g., "atgcatt") should be enclosed in quotation marks. Literal sequences are distinguished from free text by context. Qualifiers that take free text as their values do not take literal sequences, and vice versa.

The `/label=' qualifier takes a feature label as its qualifier. Although feature labels are optional, they allow unambiguous

references to the feature. The feature label identifies a feature within an entry; when combined with the accession number and the name of the data bank from which it came, it is a unique tag for that feature. Feature labels must be unique within an entry, but can be the same as a feature label in another entry. Feature labels are not case sensitive; they can be entered in any combination of upper-and lowercase without changing their meaning.

The following is a list of valid feature qualifiers.

/anticodon Location of the anticodon of tRNA and the amino acid for which it codes

/bound_moiety Moiety bound

/citation Reference to a citation providing the claim of or evidence for a feature

/codon Specifies a codon that is different from any found in the reference genetic code

/codon_start Indicates the first base of the first complete codon in a CDS (as 1 or 2 or 3)

/cons_splice Identifies intron splice sites that do not conform to the 5'-GT... AG-3' splice site consensus

/db_xref A database cross-reference; pointer to related information in another database

/direction Direction of DNA replication

/EC_number Enzyme Commission number for the enzyme product of the sequence

/evidence Value indicating the nature of supporting evidence

/frequency Frequency of the occurrence of a feature

/function Function attributed to a sequence

/gene Symbol of the gene corresponding to a sequence region (usable with all features)

/label A label used to permanently identify a feature

/map Map position of the feature in free-format text

/mod_base Abbreviation for a modified nucleotide base

/note Any comment or additional information

/number A number indicating the order of genetic elements (e.g., exons or introns) in the 5 to 3 direction

/organism Name of organism if different from that contained in the entry's ORGANISM field

/partial Differentiates between complete regions and partial ones

/phenotype Phenotype conferred by the feature

/product Name of a product encoded by the sequence

/pseudo Indicates that this feature is a non-functional version of the element named by the feature key

/rpt_family Type of repeated sequence; Alu or Kpn, for example

/rpt_type Organization of repeated sequence

/rpt_unit Identity of repeat unit that constitutes a repeat_region

/standard_name Accepted standard name for this feature

/transl_except Translational exception: single codon, the translation of which does not conform to the reference genetic code

/translation Amino acid translation of coding region (automatically generated)

/type Name of a strain if different from that in the SOURCE field

/usedin Indicates that feature is used in a compound feature in another entry

3.5.12.4 Cross-Reference Information

One type of information in the feature table lists cross-references to the annual compilation of transfer RNA sequences in Nucleic Acids Research, which has kindly been sent to us on CD-ROM by Dr. Sprinzl. Each tRNA entry of the feature table contains a /note= qualifier that includes a reference such as `(NAR: 1234)' to identify code 1234 in the NAR compilation. When such a cross-reference appears in an entry that contains a gene coding for a transfer RNA molecule, it refers to the code in the tRNA gene compilation. Similar cross-references in entries containing mature transfer RNA sequences refer to the companion compilation of tRNA sequences published by D.H. Gauss and M. Sprinzl in Nucleic Acids Research.

3.5.12.5 Feature Table Examples

In the first example a number of key names, feature locations, and qualifiers are illustrated, taken from different sequences. The first table entry is a coding region consisting of a simple span of bases and including a /gene qualifier. In the second table entry, an NAR cross-reference is given (see the previous section for a discussion of these cross-references). The third and fourth table entries use the symbols `<`and `>' to indicate that the beginning or end of the feature is beyond the range of the presented sequence. In the fifth table entry, the symbol `^' indicates that the feature is between bases.

```
1       10        20        30        40        50        60        70        79
--------+---------+---------+---------+---------+---------+---------+---------
     CDS             5..1261
                     /product="alpha-1-antitrypsin precursor"
                     /map="14q32.1"
                     /gene="PI"
     tRNA            1..87
```

```
                    /note="Leu-tRNA-CAA (NAR: 1057)"
                    /anticodon=(pos:35..37,aa:Leu)
     mRNA           1..>66
                    /note="alpha-1-acid glycoprotein mRNA"
     transposon     <1..267
                    /note="insertion element IS5"
     misc_recomb    105^106
                    /note="B.subtilis DNA end/IS5 DNA start"
     conflict       258
                    /replace="t"
                    /citation=[2]
---------+---------+---------+---------+---------+---------+---------+---------
1        10        20        30        40        50        60        70        79
```

Example 10. Feature Table Entries

The next example shows the representation for a CDS that spans more than one entry.

```
1        10        20        30        40        50        60        70        79
---------+---------+---------+---------+---------+---------+---------+---------
LOCUS       HUMPGAMM1    3688 bp ds-DNA             PRI       15-OCT-1990
DEFINITION  Human phosphoglycerate mutase (muscle specific isozyme) (PGAM-M)
            gene, 5' end.
ACCESSION   M55673 M25818 M27095
KEYWORDS    phosphoglycerate mutase.
SEGMENT     1 of 2
  .
  .
  .
FEATURES             Location/Qualifiers
     CAAT_signal     1751..1755
                     /gene="PGAM-M"
     TATA_signal     1791..1799
                     /gene="PGAM-M"
     exon            1820..2274
                     /number=1
                     /EC_number="5.4.2.1"
                     /gene="PGAM-M"
     intron          2275..2377
                     /number=1
                     /gene="PGAM2"
     exon            2378..2558
                     /number=2
                     /gene="PGAM-M"
  .
  .
  .
//
LOCUS       HUMPGAMM2     677 bp ds-DNA             PRI       15-OCT-1990
DEFINITION  Human phosphoglycerate mutase (muscle specific isozyme) (PGAM-M),
            exon 3.
ACCESSION   M55674 M25818 M27096
KEYWORDS    phosphoglycerate mutase.
SEGMENT     2 of 2
  .
  .
  .
FEATURES             Location/Qualifiers
```

```
     exon            255..457
                     /number=3
                     /gene="PGAM-M"
     intron          order(M55673:2559..>3688,<1..254)
                     /number=2
                     /gene="PGAM-M"
     mRNA            join(M55673:1820..2274,M55673:2378..2558,255..457)
                     /gene="PGAM-M"
     CDS             join(M55673:1861..2274,M55673:2378..2558,255..421)
                     /note="muscle-specific isozyme"
                     /gene="PGAM2"
                     /product="phosphoglycerate mutase"
                     /codon_start=1
                     /translation="MATHRLVMVRHGESTWNQENRFCGWFDAELSEKGTEEAKRGAKA
                     IKDAKMEFDICYTSVLKRAIRTLWAILDGTDQMWLPVVRTWRLNERHYGGLTGLNKAE
                     TAAKHGEEQVKIWRRSFDIPPPPMDEKHPYYNSISKERRYAGLKPGELPTCESLKDTI
                     ARALPFWNEEIVPQIKAGKRVLIAAHGNSLRGIVKHLEGMSDQAIMELNLPTGIPIVY
                     ELNKELKPTKPMQFLGDEETVRKAMEAVAAQGKAK"
   .
   .
   .
//
---------+---------+---------+---------+---------+---------+---------+---------
1        10        20        30        40        50        60        70        79
```

Example 11. Joining Sequences

3.5.13 ORIGIN Format

The ORIGIN record may be left blank, may appear as `Unreported.' or may give a local pointer to the sequence start, usually involving an experimentally determined restriction cleavage site or the genetic locus (if available). The ORIGIN record ends in a period if it contains data, but does not include the period if the record is left empty (in contrast to the KEYWORDS field which contains a period rather than being left blank).

3.5.14 SEQUENCE Format

The nucleotide sequence for an entry is found in the records following the ORIGIN record. The sequence is reported in the 5 to 3 direction. There are sixty bases per record, listed in groups of ten bases followed by a blank, starting at position 11 of each record. The number of the first nucleotide in the record is given in columns 4 to 9 (right justified) of the record.

4. ALTERNATE RELEASES

NCBI is supplying sequence data in the GenBank flat file format to maintain compatibility with existing software which require that particular format. Although we have made every effort to ensure that these data are presented in the traditional flat file format, if you encounter any problems in using these data with software which is based upon the flat file format, please contact us at:

info@ncbi.nlm.nih.gov

The flat file is just one of many possible report formats that can be

generated from the richer representation supported by the ASN.1 form of the data. Developers of new software tools should consider using the ASN.1 form directly to take advantage of those features. Documentation and a Software Developer's Toolkit for ASN.1 are available through NCBI. You may call NCBI at (301)496-2475, or subscribe to a developers' electronic newsgroup by sending your name, address, affiliation, and e-mail address to:

bits-request@ncbi.nlm.nih.gov

The Software Developer's Toolkit and PostScript documentation for UNIX, VMS, Ultrix, AIX, MacOS, DOS, and Microsoft Windows systems is available in a compressed UNIX tar file by anonymous ftp from 'ncbi.nlm.nih.gov', in the toolbox/ncbi_tools directory. The file is 'ncbi.tar.Z'.

5. KNOWN PROBLEMS OF THE GENBANK DATABASE

5.1 Incorrect Gene Symbols in Entries and Index

The /gene qualifier for many GenBank entries contains values other than the official gene symbol, such as the product or the standard name of the gene. The gene symbol index (gbgen.idx) is created from the data in the /gene qualifier and therefore may contain data other than official gene symbols.

6. GENBANK ADMINISTRATION

The National Center for Biotechnology Information (NCBI), National Library of Medicine, National Institutes of Health, is responsible for the production and distribution of the NIH GenBank Sequence Database. NCBI distributes GenBank sequence data by CD-ROM, anonymous FTP, e-mail servers and other network services. For more information, you may contact NCBI at the e-mail address: info@ncbi.nlm.nih.gov or by phone: 301-496-2475.

6.1 Registered Trademark Notice

GenBank (R) is a registered trademark of the U.S. Department of Health and Human Services for the Genetic Sequence Data Bank.

6.2 Citing GenBank

If you have used GenBank in your research, we would appreciate it if you would include a reference to GenBank in all publications related to that research.

When citing data in GenBank, it is appropriate to give the sequence name, primary accession number, and the publication in which the sequence first appeared. If the data are unpublished, we urge you to contact the group which submitted the data to GenBank to see if there is a recent publication or if they have determined any revisions or extensions of the data.

It is also appropriate to list a reference for GenBank itself. The following publication, which describes the GenBank database, should be cited:

Benson, D.A., Boguski, M.S., Lipman, D.J., and Ostell, J.
GenBank. Nucl. Acids Res. 25(1):1-6 (1997)

The following statement is an example of how you may cite GenBank

data. It cites the sequence, its primary accession number, the group who determined the sequence, and GenBank. The numbers in parentheses refer to the GenBank citation above and to the REFERENCE in the GenBank sequence entry.

`We scanned the GenBank (1) data bank for sequence similarities and found one sequence (2), GenBank accession number J01016, which showed significant similarity...'

(1) Benson, D.A. et al. Nucl. Acids Res. 25(1):1-6 (1997)
(2) Nellen, W. and Gallwitz, D. J. Mol. Biol. 159, 1-18 (1982)

6.3 GenBank Distribution Formats and Media

GenBank data are available on industry-standard ISO-9660 CD-ROM. The standard flat file format is included.

This documentation accompanies the ten CD-ROM set entitled 'GenBank (Flat File Format).' Each release is cumulative, incorporating all previous GenBank data. No retrieval software is provided.

6.4 Other CD-ROM Titles

The Entrez CDROM release has been discontined, effective August 15, 1996.

Entrez is a molecular biology database system that presents an integrated view of DNA and protein sequence data, 3D structure data, and associated MEDLINE entries. The system is produced by the National Center for Biotechnology Information (NCBI), and is now available only over the Internet.

The CD-ROM version was discontinued due to the continuing rapid growth in size of GenBank and other sequence databases, causing the CD-ROM release to become increasingly unwieldy and inconvenient. The CD-ROM version also lagged far behind the two Internet versions, Network Entrez and Web Entrez, in the number of MEDLINE citations available and in the addition of new databases, including the Genomes division and the Structure division, as well as links to an increasing number of on-line journals. In addition, the two on-line versions of Entrez are updated daily, compared to the bimonthly CD-ROM updates.

Access to the Internet versions of Entrez is easy. If you have a World Wide Web browser, such as Netscape or Explorer, simply point your browser to http://www.ncbi.nlm.nih.gov/. The Web version of Entrez has all the capabilities of the CD-ROM version, but with the visual style of the World Wide Web. If you preferred the "look and feel" of the CD-ROM version, you may download Network Entrez from the NCBI's anonymous FTP site: ncbi.nlm.nih.gov. Versions are available for PC/Windows, Macintosh and several Unix workstations.

For information about Network Entrez, Web Entrez or any other NCBI services, you may contact NCBI by e-mail to info@ncbi.nlm.nih.gov or by phone at 301-496-2475.

Ordering Information

GPO will handle all subscriptions and subscription-related questions. Telephone orders can be placed at (202) 512-1800. Due to the high volume of telephone ordering, GPO encourages ordering by fax or mail. Quantity discounts of 25% are available for orders of 100 or more CD-ROMs delivered to a single address. We strongly recommend that you use a check, money order

or credit card to order your CD-ROM. If you absolutely must use a purchase order, you will need prior approval from the Government Printing Office (GPO). To obtain the necessary approval, call GPO at (202)512-2477 and request "Fill and Bill Application Form #3695".

==========

Superintendent of Documents CD-ROM Subscription Order Form

Order processing code:
7545

To fax your orders:
(202)512-2233
To phone your orders:
(202)512-1800

|_| YES, enter my order as follows:

___ subscriptions to GenBank (Flat File) on CD-ROM (NCBIF)

The total cost of my order is $_____. Price includes postage and handling and is subject to change.

_____
(Company or personal name)

_____
(Additional address/attention line)

_____
(Street address)

_____
(City, State, ZIP code)

_____
(Daytime phone including area code)

_____
(Purchase Order No. for pre-approved "Fill and Bill" customers only)

To open a fill and bill acct.,your billing office must complete Fill and Bill Application Form #3695. To obtain a fill and bill application form call (202)512-2477.

For privacy protection, check the box below:
|_| Do not make my name available to other mailers.

Please choose method of payment:

___ Check Payable to the Superintendent of Documents
___ GPO Deposit Account
|_|_|_|_|_|_|_| - |_|
___ VISA or MasterCard Account

|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|_|

|_|_|_|_| (Credit card expiration date)

_____
(Authorizing signature) 10/94

Mail To: Superintendent of Documents
P.O. Box 371954
Pittsburgh, PA 15250-7954

One Year Subscription Prices (six issues) Including Delivery

Location	GenBank
United States	$192.00
Zone 1 (South America)	$240.00
Zone 2 (Europe)	$240.00

Zone 3 (Middle East and Africa).....$240.00
Zone 4 (Asia, the Pacific, Australia).....$240.00
Canada.....$240.00
Mexico.....$240.00

U.S. delivery via first class mail. Foreign delivery via air mail. Single copy price of the GenBank CDROM is $40.00 in the United States, and $50.00 outside the United States.

6.5 Request for Corrections and Comments

We welcome your suggestions for improvements to GenBank. We are especially interested to learn of errors or inconsistencies in the data. Please use the GenBank Error/Suggestion Report Form, which is part of this distribution of GenBank (located in the file gbdat.frm), to send your suggestions and corrections by electronic mail to: update@ncbi.nlm.nih.gov or to the address on the error/suggestion form. Please be certain to indicate the GenBank release number (e.g., Release 104.0) and the primary accession number of the entry to which your comments apply; it is helpful if you also give the entry name and the current contents of any data field for which you are recommending a change.

6.6 Credits and Acknowledgments

Credits -

CD-ROM Data Coordination
Mark Cavanaugh

GenBank Submission Coordination
Francis Ouellette

GenBank Annotation Staff
Kirsty Bates, Karen Clark, Irene Fang, Michael Fetchko, Pamela Jacques, Junga Kim, Maureen Madden, Simos Marmaras, Baishali Maskeri, Jenny McDowell, Ilene Mizrachi, Steve Parker, Joseph Plante, Lillian Riddick, Leigh Riley, and Dorothy Trinh, Jane Weisemann

Data Management and Preparation
Mark Cavanaugh, Tatiana Tatusov, Hsiu-Chuan Chen, Jane Weisemann, Carolyn Tolstoshev, Karl Sirotkin, Jim Ostell

Database Administration
Eugene Yaschenko, Cat Lukens

Production and User Support
Dennis Benson, Terri Clark, Jim Fleshman, Renata McCarthy, Scott McGinnis, Monica Romiti, Barbara Rapp, Steven Rosenthal, Rose Marie Woodsmall

Graphics
Greg Schuler

Project Direction
David Lipman

Acknowledgments -

Contractor support for GenBank production and distribution
has been provided by Management Systems Designers, Inc.,
ComputerCraft Corporation, and The KEVRIC Company, Inc.

6.7 Disclaimer

The United States Government makes no representations or warranties regarding the content or accuracy of the information. The United States Government also makes no representations or warranties of merchantability or fitness for a particular purpose and accepts no responsibility for any consequences of the receipt or use of the information.

For additional information about NCBI distributions, please contact NCBI by e-mail at info@ncbi.nlm.nih.gov, by phone at (301) 496-2475, or by mail at:

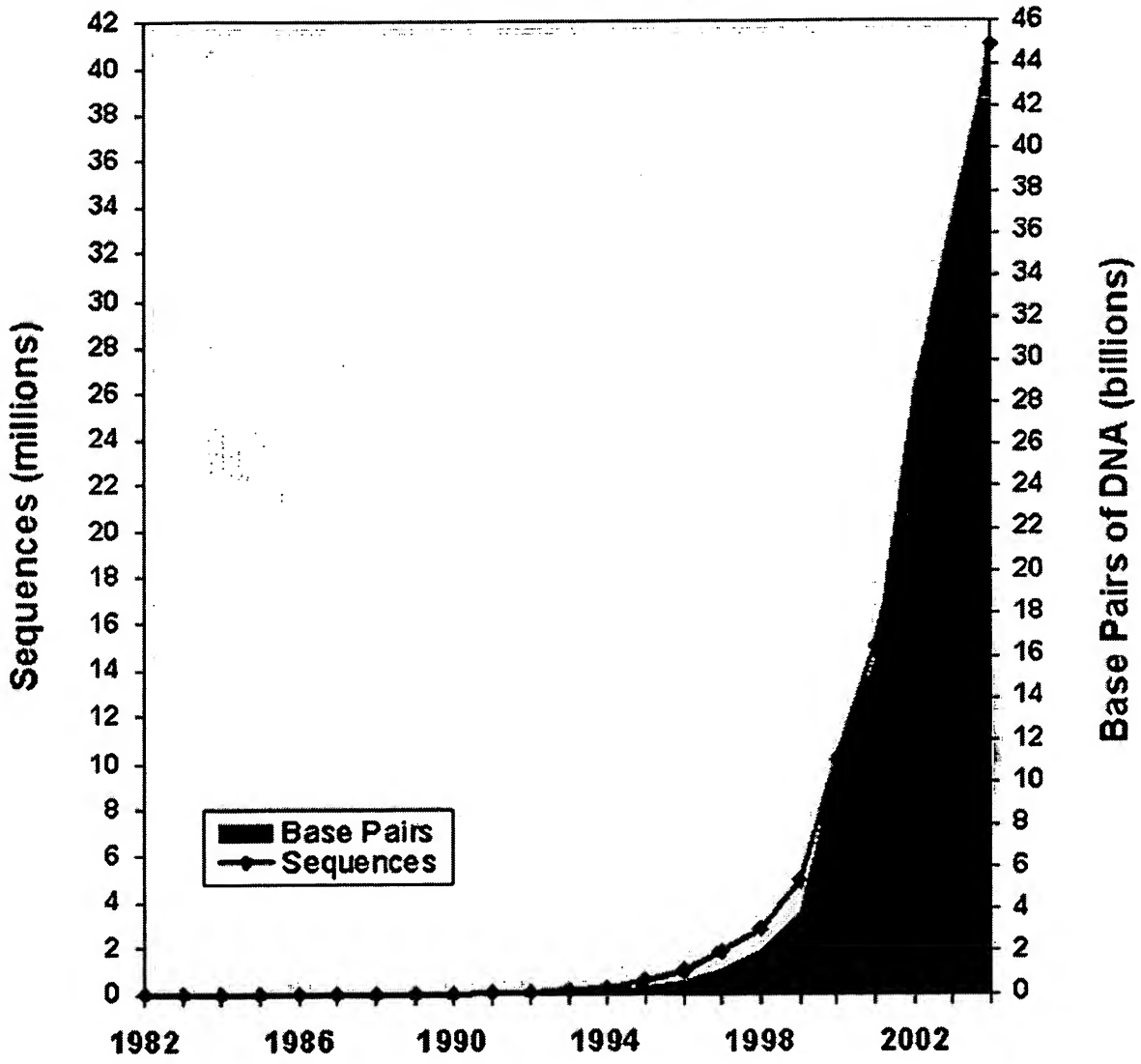
GenBank
National Library of Medicine
Bldg. 38A Rm. 8N-809
8600 Rockville Pike
Bethesda, MD 20894
FAX: (301) 480-9241

NCBI **GenBank Statistics**

PubMed Entrez BLAST OMIM Books Taxonomy Structure

Growth of GenBank

(1982 - 2004)



Release Notes for the current verison of GenBank

GenBank Data		
Year	Base Pairs	Sequences

1982	680,338	606
1983	2,274,029	2,427
1984	3,368,765	4,175
1985	5,204,420	5,700
1986	9,615,371	9,978
1987	15,514,776	14,584
1988	23,800,000	20,579
1989	34,762,585	28,791
1990	49,179,285	39,533
1991	71,947,426	55,627
1992	101,008,486	78,608
1993	157,152,442	143,492
1994	217,102,462	215,273
1995	384,939,485	555,694
1996	651,972,984	1,021,211
1997	1,160,300,687	1,765,847
1998	2,008,761,784	2,837,897
1999	3,841,163,011	4,864,570
2000	11,101,066,288	10,106,023
2001	15,849,921,438	14,976,310
2002	28,507,990,166	22,318,883
2003	36,553,368,485	30,968,418
2004	44,575,745,176	40,604,319

Revised: February 16, 2005.

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicant	:	Remacle, et al.
Appl. No.	:	10/035,822
Filed	:	December 27, 2001
For	:	DETECTION AND/OR QUANTIFICATION OF A TARGET MOLECULE BY BINDING WITH A CAPTURE MOLECULE FIXED ON THE SURFACE OF A DISC
Examiner	:	Sisson, Bradley L.
Group Art Unit:		1634

DECLARATION UNDER 37 C.F.R. §1.132

United States Patent and Trademark Office
P.O. Box 2327
Arlington, VA 22202

Dear Sir:

1. This Declaration is being submitted to demonstrate that at as of December 30, 1997 (the date we filed US Provisional Patent Application No. 60/071,726, to which the present application claims priority) one of skill in the art would recognize that the methodology described in the specification for binding a capture molecule to the disc can be used with any desired capture molecule, that as of that date one of skill in the art was familiar with technology for converting a signal from the claimed discs into a desired form of output (such as words, numbers, notes etc), that as of that date one of skill in the art was familiar with reactants which can be used for binding a target molecule in a sample to a capture molecule on the claimed discs and for detecting the binding of a target molecule to a capture molecule on the surface of the claimed discs, that as of that date one of skill in the art was familiar with the use of lasers to read data

from a disc, and that as of that date one of skill in the art was familiar with technology which can be used for quantitation of a signal on the claimed discs.

2. I am an inventor on the above-identified patent application and am familiar with the specification and prosecution history.

3. I have extensive experience in the field of the claimed invention as indicated in the attached Curriculum Vitae provided as Exhibit A of the Declaration.

4. The claimed invention relates to discs comprising registered data, and bound to its surface one or more non-cleavable capture molecules which allow for binding with one or more target molecules to be detected, identified and/or quantified. In some embodiments, the capture molecules are nucleotide sequences. In other embodiments, the capture molecules are peptides or polypeptides, such as antibodies, receptors, and enzymes. In other embodiments, the capture molecules are antigens or ligands of receptors, which in some instances may also be polypeptides or peptides. In other embodiments, the capture molecules may be lipids, saccharides, haptens, fluorophores, chromophores, catalysts, new macromolecules obtained by combinatorial chemistry. In further embodiments, the capture molecules may be combinations of any of the foregoing.

5. With respect to embodiments in which the capture molecule is a nucleic acid, the surface of the disc may be aminated as described, for example, on page 19, lines 15-31 of the Provisional application, as well as on page 20, lines 16-18, page 44, line 31-page 45, line 9, and in Example 1 of the present specification, thereby the nucleic acids may be bound to the amine groups on the disc as described at the foregoing locations of the Provisional and the present applications. As of December 30, 1997, those of skill in the art would appreciate that because the amine groups on the surface of the disc can be covalently bound to any nucleic acid regardless of its sequence, the methodology described in the specification is universally applicable to all nucleic acids. Accordingly, as of December 30, 1997 those skilled in the art would appreciate that the application contained sufficient description of how to bind any desired nucleic acid to the surface of the disc.

6. With respect to embodiments in which the capture molecule is a peptide or polypeptide, the surface of the disc may be carboxylated as described, for example, on page 19, lines 15-31 of the Provisional application, as well as in Examples 3-5 of the present application and the peptide or polypeptide can be bound to carboxyl groups as described at the foregoing locations in the specifications. As of December 30, 1997, those of skill in the art would appreciate that because the carboxyl groups on the surface of the disc can be covalently bound to any peptide or polypeptide regardless of its amino acid sequence, the methodology described in the specification is universally applicable to all peptides or polypeptides. Accordingly, as of December 30, 1997 those skilled in the art would appreciate that the application contained sufficient description of how to bind any desired peptide or polypeptide to the surface of the disc.

7. Other methodology for fixing capture molecules bearing amino groups to the surface of the disc is described in the present specification at page 20, lines 7-16. As of December 30, 1997, those of skill in the art would appreciate that because this methodology will work with any capture molecule bearing an amino group, the application contained sufficient description of how to bind any desired capture molecule bearing an amino group to the surface of the disc (see, for example, page 19, lines 15-31 of the Provisional application and Rasmussen et al. 1991 "Covalent immobilization of DNA onto polystyrene microwells: the molecules are only bound at the 5" end" Anal. Biochem. 198:138-142, which is referenced in the Provisional application on page 24, line 32). This methodology would work for haptens, fluorophores, chromophores, catalysts, new macromolecules obtained by combinatorial chemistry, or any other molecule bearing an amino group. Thus, as of December 30, 1997 those skilled in the art would appreciate that the application contained sufficient description of how to bind any desired molecule bearing an amino group to the surface of the disc.

8. Other methods for fixing any desired capture molecule bearing a reactive group by deprotecting or protecting the reactive group or by synthesizing the capture molecule on the surface of the disc are described in the present specification at page 30, line 19 - page 31, line 2.

9. As of December 30, 1997, those skilled in the art were familiar with how to convert digital information on the disc into a desired form of output, such as words, notes, numbers etc. As described in the Provisional application at page 10, lines 13-32, and in the present

specification at page 4, lines 1-35, and in the conventional CD technology available on December 30, 1997, data is stored on the CD as pits and lands. The pits and lands are converted into digital data (1's and 0's) when read by a laser. Specifically, each pit is converted into a binary 1 and each land is converted into a binary 0. As of December 30, 1997, CD's were being utilized to provide output in a variety of formats and those skilled in the art would appreciate that such technology was standard. In fact, as demonstrated in the attached History of CD Technology (Exhibit 1) CD technology was quite mature at the time the present application was filed. Thus, as of December 30, 1997, those skilled in the art could readily convert binary digital information into any desired form of output.

10. Similarly, as of December 30, 1997, those skilled in the art were familiar with the use of lasers to read data from a disc (see Provisional application at page 9, line 13 through page 10, line 25, page 11, line 31 - through page 12, line 29). As discussed in the preceding paragraph, data present on the disc is converted into 1's or 0's using conventional laser technology. The presently claimed discs contain registered data stored on the disc as conventional pits or lands. When the laser light shines on a pit, a signal transition is generated which is converted into a binary 1 (see present specification at page 11, lines 17-31). Likewise, when the laser light shines on a land, no signal transition is generated and this is converted into a binary 0. Again, this technology is the standard technology which was conventionally used as of December 30, 1997 to retrieve data from a standard CD at the time the present application was filed.

The presently claimed discs also generate binary data reflecting the binding of a target to a capture molecule. In some embodiments, binding of a target to a capture molecule results in formation of a precipitate, which forms a mound on the surface of the disc. The mound perturbs the laser reflection (just as a pit perturbs the laser reflection) and is converted into a binary 1 (see Provisional application, page 14, line 18 through page 16, line 24, page 21, lines 20-23, and Figure 3; and the present specification, page 23, line 24-page 24, line 33) Thus, binding of a target to a capture molecule is detected using the standard laser technology used to read data from a conventional CD which was available as of December 30, 1997.

11. As of December 30, 1997, those skilled in the art were familiar with reactants which can be used for binding a target molecule in a sample to a capture molecule on the discs of the

present invention. In particular, as of December 30, 1997 those skilled in the art were familiar with appropriate buffers and binding conditions suitable for conducting nucleic acid hybridizations, ligand/polypeptide binding reactions, antibody/antigen binding reactions, lipid binding reactions, and saccharide binding reactions (see Provisional application, page 1, line 25 - page 4, line 12 and page 20, lines 9-30). In addition, particular buffers and conditions are set forth in the present application at Examples 1-5, Example 7-10 and Example 16 of the specification. In addition, as of December 30, 1997 a variety of DNA chips were available and those skilled in the art were familiar with buffers and hybridization conditions which could be used with such chips (see "DNA chips: State-of-the-art", published in Jan. 1998 and discussing various DNA chip systems, provided herewith as Exhibit 2)

In addition, as of December 30, 1997, those skilled in the art were familiar with reactants which can be used to detect binding between a target molecule and a sample. In particular, a variety of reactants which could be used to form precipitates at a bound target were known as of December 30, 1997. For example, reactants for generating a precipitate by silver staining, fluorescent reagents, and colorimetric reactants were known to those of skill as of December 30, 1997 (see Provisional application at page 3, line 30 - page 4, line 6, and page 22, line 9 - page 24, line 7). In addition, the present specification provides numerous examples of such reactants at page 17, line 28-page 18, line 26, page 22, line 4-page 25, line 8 and Examples 1-10, and Examples 13-15.

12. As of December 30, 1997, those of skill in the art were familiar with technology which can be used for quantitating the signal resulting from binding of a target to a capture molecule on the surface of the claimed discs. For example, binding of a target to a capture molecule can be quantitated using standard laser technology employed in conventional CD players which were available as of December 30, 1997, standard fluorescence reading technologies available as of December 30, 1997, and image recognition software available as of December 30, 1997 (see Provisional application, page 3, line 6 - page 4, line 12, and page 24, lines 8-15).

The present specification also describes such technologies In addition, the present specification describes quantitation at pages 18-23, page 45, lines 10-2. In addition, Figures 17 and 20 provide actual quantitation curves obtained with the presently claimed discs.

13. I declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful, false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or patent issuing therefrom.

Dated: 1 March 2005

By: _____
Jose Remacle